Testing*

Annika B. Bergbauer, Eric A. Hanushek, and Ludger Woessmann†

Abstract

The significant expansion of varying forms of student testing, while controversial in many countries, has not been generally linked to educational performance. Different testing regimes provide varying information to parents, teachers, and decision makers. We investigate how different types of information relate to student achievement. Our empirical analysis exploits data from over two million students in 59 countries observed across six waves of the international PISA test 2000-2015. Relying on the country panel feature of the data, we investigate how testing reforms relate to country performance on PISA tests over time, taking out country and year fixed effects. Expansion of standardized testing with external comparison, both school-based and student-based, is associated with improvements in student achievement. This effect is stronger in low-performing countries. By contrast, reforms to solely internal testing without external comparison and internal teacher monitoring including inspectorates are not related to changes in student achievement.

Keywords: student assessment, testing, student achievement, international, PISA

JEL classification: I28, H52, L15, D82, P51

November 7, 2019

* We gratefully acknowledge comments from Scott Imberman, Joachim Winter, and participants at seminars in Berlin, Maastricht, Madrid, and Moscow, the European Association of Labour Economists in Lyon, the Association for Education Finance and Policy in Kansas City, the Spring Meeting of Young Economists in Palma de Mallorca, the German Economic Association in Freiburg, the CESifo Area Conference on Economics of Education in Munich, the briq Workshop Skills, Preferences and Educational Inequality in Bonn, the CRC 190 meeting in Ohlstadt, the BGPE Research Workshop in Bamberg, and the Center Seminar of the ifo Center for the Economics of Education in Munich. This work was supported by the Smith Richardson Foundation. The contribution by Bergbauer and Woessmann is part of project CRC TRR 190 of the German Science Foundation. Woessmann acknowledges support by Research on Improving Systems of Education (RISE) which is funded by UK Aid and Australian Aid.

† Bergbauer: ifo Institute at the University of Munich, bergbauer@ifo.de; Hanushek: Hoover Institution, Stanford University, CESifo, IZA, and NBER, hanushek@stanford.edu; Woessmann: University of Munich, ifo Institute, CESifo, and IZA, woessmann@ifo.de.

1. Introduction

Student testing has grown rapidly around the world. While some have argued that this trend has been damaging to schooling (Hout and Elliott (2011); Andrews and coauthors (2014)), others have argued that even more testing is called for. In fact, the World Bank (2018), in evaluating the need for improved human capital development around the world, explicitly calls for expansion of student evaluations and concludes that "[t]here is too little measurement of learning, not too much" (p. 17). However, both critics and proponents of international and national testing often fail to differentiate among alternative forms of testing, leading to a confused debate.

Understanding the overall impact of student testing requires careful consideration of a test's specific informational content, which determines how the assessments can be used. In the United States, for example, consideration of testing is mostly restricted to the specific accountability systems exemplified by No Child Left Behind (NCLB), a 2001 federal law that required states to test student outcomes annually in grades 3-8 and intervene in schools that were not on track to bring all students to state-defined proficiency levels. But testing students to assess the performance of schools is very different from evaluating teachers based on student performance or from using tests to select which students should continue on to university. And standardized tests normed to a large population are very different from teacher-generated tests used to assess the pace of classroom learning or from monitoring teacher practices in classrooms locally.

This paper exploits international comparisons to examine how different types of testing contribute to overall levels of student achievement. We argue that student assessments (used as a synonym for testing here) provide to varying degrees the informational backbone for alternative policies and incentive systems that can lead to various behavioral results. Based on the conceptual framework of a principal-agent model, we discuss the kind of information created by

a continuum of forms of testing from internal assessments to standardized external comparisons. We are interested in the reduced-form effect of the availability of testing per se, rather than how the generated information is used in any particular policies or accountability systems.

Our empirical analysis uses data from the Programme for International Student Assessment (PISA) to construct a panel of country observations of student performance. Specifically, we pool the micro data of over two million students across 59 countries participating in six PISA waves between 2000 and 2015. PISA includes not only measures of student outcomes, but also rich background information on both students and schooling institutions in the different countries. From these surveys and other international data sources, we also derive a series of measures of different types of student testing that allow us to differentiate four testing categories: at the two ends of the continuum, (1) internal testing and (2) standardized testing with external comparison; in-between, (3) standardized testing without external comparison; and, while generally abstracting from particular uses, we separate out (4) internal teacher monitoring from other forms of internal testing as the one category that cannot be separated from its specific use.

Because this is a period of rapid change in student assessment policies across countries, we can link information policies to outcomes in fixed-effects panel models. Our identification relies on changes in student assessment regimes within countries over time. While using the individual student data for estimation at the micro level, we measure our treatment variables as country aggregates at each point in time to avoid bias from within-country selection of students into schools. Conditioning on country and year fixed effects allows us to account for unobserved time-invariant country characteristics as well as common time-specific shocks.[1]

[1] Our analysis expands on the growing literature studying determinants of student achievement in a cross-country setting (Hanushek and Woessmann (2011); Woessmann (2016)). Methodologically, our approach builds on the analysis of school autonomy in Hanushek, Link, and Woessmann (2013).

Our analysis shows that some forms of student testing are strongly related to student learning while others have no discernible association with learning outcomes. On the one hand, expansion of standardized testing that provides external comparisons is associated with increased performance on the international tests. This is true for both school-based and student-based forms of external comparisons and in math, science, and reading. On the other hand, internal testing that simply informs or monitors progress without external comparability and internal teacher monitoring including inspectorates have little discernible effect on overall performance. Standardized testing without external comparison, while not related to student achievement on average, has a positive effect in initially poorly performing countries but not in initially highly performing countries. Similarly, the impact of school-based external comparisons differs across schooling systems with larger impacts in poorer performing systems.

In a placebo test with leads of the testing variables, we show that changes in assessment are not systematically linked to prior outcome conditions. We also discuss a number of specification tests that speak against substantial bias from coincidental other policies and show that results hold in a long-difference specification. Furthermore, robustness tests show that results are not affected by any individual country, by consideration of subsets of countries, by controlling for test exclusion rates, by changes in PISA testing procedures, and by estimating the model collapsed to the country-by-wave level.

Our cross-country approach allows us to draw on the substantial variation in different forms of testing that exists across countries. Testing policies are often set at the national level, making it difficult to construct an adequate comparison group for evaluation of any outcomes in a within-country setting. By moving to international comparisons, it is possible to study these national policies, investigate which aspects of student assessment systems generalize to larger

settings and which do not, and consider how overall institutional structures interact with the specifics of student assessment systems. Of course, these advantages come at a cost, because precisely identifying the separate impact of information across nations offers its own challenges. We are not able to investigate the details of specific national programs and policies that might rely on the information created, and there is uncertainty in separating the changes in information flows from the range of individual programs, policies, and usages developed from them. Through a variety of approaches, our additional analyses can reduce concerns of substantial bias from the most obvious sources in the cross-country setting, but they cannot completely eliminate any possible biases. However, only the comparative perspective allows for an investigation of the richness of the full continuum of different forms of testing by exploiting the counterfactual from countries that did not reform at the same point in time.

In the literature (discussed within our conceptual framework in section 2), as well as in policy discussions, the term testing is frequently taken to be synonymous with accountability. We think it is useful to separate these two concepts. Accountability systems link various learning outcomes to rewards, punishments, and incentives for different actors, and they can differ widely in form and substance. Moreover, any given student assessment can simultaneously be used in multiple ways for accountability purposes. Testing also enters into educational decision making in broader ways than just accountability. Information from student assessments is used in policy formulation, program evaluations, and regulatory structures. We therefore think of the various student assessments as providing information necessary for implementing different sets of policies, potentially inducing behavioral changes that affect learning outcomes. From a policy perspective, a focus on testing is useful because policy makers cannot always fully control how information is used by different actors, but they can in general institute which type of testing

information is provided. We therefore consider our study as a reduced-form analysis that focuses on how the informational content of different testing regimes can support policies, programs, and actions that lead to altered student outcomes but does not delve into the structures of any specific policies or accountability systems that are subsequently attached to the assessment.

The next section develops a conceptual framework highlighting different forms of student assessments. Section 3 introduces the data and Section 4 the empirical model. Section 5 presents our results including analyses of heterogeneous effects. Section 6 reports a placebo test and other specification tests, and Section 7 shows a series of robustness analyses. Section 8 concludes.

2. Evaluating Testing

We begin with the conceptual framework of a principal-agent structure and identify the continuum of internal to external forms of testing as a key aspect that motivates our empirical modeling.[2]

2.1 The Principal-Agent Framework

A useful way to characterize the structure of educational systems is as a tree of principal-agent problems (Laffont and Martimort (2002)).[3] Parents care about their child's achievement of knowledge and skills, which directly affects their long-run economic outcomes (Card (1999); Hanushek, Schwerdt, Wiederhold, and Woessmann (2015)). Parents, however, cannot directly choose the effort level of their children. Instead, they may offer short-term rewards for learning to their child and try as best as possible to observe and control child effort. Similarly, parents

[2] For a more extensive discussion of the conceptual framework that covers the underlying value functions and the technology of student assessment in greater detail, see the working-paper version of this paper (Bergbauer, Hanushek, and Woessmann (2018)).

[3] See Bishop and Woessmann (2004) and Pritchett (2015) for related analyses of education systems as principal-agent relationships.

cannot fully control the production of the child's achievement in schools, where a key element is the effort levels of teachers and other school personnel.

Parents act as principals that contract the teaching of their children to schools and teachers as agents. In the process of classroom instruction, teachers also act as principals themselves who cannot fully observe the learning effort of their students as agents. Teaching in the classroom and studying at a desk involve asymmetric information in that the respective principal cannot fully monitor the behavior of the respective agent. Because of the incomplete monitoring and the specific objective functions of parents, teachers, and students, one cannot simply assume that the actions of children and teachers will lead to the optimal result for parents.

In addition to the parent-child problem, the parent-teacher problem, and the teacher-child problem as canonical elements of the tree of principal-agent relationships, the administration adds another layer to the system. Parents often look beyond the individual teacher to school administrators at different levels, including the nation, the region, the school district, and the school. This suggests that there are parent-administrator problems, administrator-administrator problems, and administrator-teacher problems that are relevant to incentive design questions.

If parents had full information about the effort levels of students, teachers, and administrators, they could effectively contract with each to maximize their own objective function. However, actually obtaining and monitoring effort levels is generally costly, and the differing preferences may lead to suboptimal effort levels by students, teachers, and administrators from the perspective of parents.

A common solution is to introduce outside assessments of the outcomes of interest. By creating outcome information, student assessments provide a mechanism for developing better

incentives to elicit increased effort by students, teachers, and administrators, thereby ultimately raising student achievement levels to better approximate the desires of the parents.

Nonetheless, a number of issues related to the type and accuracy of information that the tests generate make the impact of testing a complicated empirical question. There is a classical identification problem of separating the joint effort levels of teachers and students in order to provide the right incentives. Additionally, imperfect measurement technologies may not provide complete information on achievement.[4] Here, we highlight that the internal vs. external character of the information generated by the test is a major source of its ability to solve the underlying principal-agent problems, with important implications for the potential impact of testing.

2.2 The Continuum from Internal Testing to Standardized External Comparison

Testing is a ubiquitous component of schooling, but not all tests create the same kind of information. By far the most common type of testing is teacher-developed tests, a form of internal testing that is used both to guide instruction and to provide feedback to students and parents. The key feature of teacher-developed tests is that their results are very difficult to compare across teachers, implying they do not provide the kind of information that would mitigate the principal-agent problem between parents and teachers. More generally, if not standardized across schools, the achievement information generated by internal testing does not directly allow parents and administrators to monitor school performance.[5] At the most extreme,

[4] Prior discussions of accountability systems have considered various dimensions of this problem (Figlio and Loeb (2011)). Perhaps the best-known conceptual discussion is the classic Holmstrom and Milgrom (1991) paper that considers how imperfect measurement of outcomes distorts incentives. In particular, if there are multiple objectives and only a subset is measured, effort could be distorted to the observed outcomes to the detriment of unobserved outcomes. But there is also more general discussion of such topics as teaching to the test (Koretz (2017)), gaming of tests (e.g., nutritious feeding on testing days, see Figlio and Winicki (2005)), and cheating (Jacob and Levitt (2003)). Each of these topics includes an element of testing technology and the accuracy of observed measures and is the subject of a much larger literature.

[5] For example, an extension of teacher-developed tests is periodic content testing provided by external producers (so-called formative assessments). Again, parents generally cannot compare outcomes externally.

tests that have no consequences for any of the actors may be inconsequential for overall performance because nobody may take them seriously.

At the other end of the continuum of testing are standardized tests that allow for external comparisons of student outcomes in different circumstances. These tests are normed to relevant population performance. The comparability of the generated achievement information suggests the possibility of using the tests to support incentives to students, but also to administrators and teachers by making external information available to parents, policy makers, and the general public.[6] As a general principle, information that is useful for producing stronger incentives is expected to have larger potential impacts on overall achievement.

The incentives created by standardized testing with external comparison may differ across the various actors, and information that helps solve one principal-agent problem may leave others untouched. In some cases, the actions of the individual actors may be plausibly separated. For example, centralized exit exams that have consequences for further schooling of students may be linked to strong incentives for student effort while having limited impact on teacher effort.[7] On the other hand, testing that is directly linked to consequences for schools such as the NCLB legislation in the US may have limited relevance for students and their efforts.[8] Similarly, differential rewards to teachers based upon test-score growth are high stakes for the teachers, but

[6] For example, school rankings may be published to the general public (see Koning and van der Wiel (2012) for the Netherlands, Burgess, Wilson, and Worth (2013) for Wales, and Nunes, Reis, and Seabra (2015) for Portugal), and school report cards may provide information to local communities (see Andrab, Das, and Khwaja (2017) for evidence from a sample of villages in Pakistan).

[7] By affecting chances to enter specific institutions and fields of higher education and the hiring decisions of employers, central exit exams usually have real consequences for students (see Bishop (1997); Woessmann (2003); Jürges, Schneider, and Büchel (2005); Woessmann, Luedemann, Schuetz, and West (2009); Luedemann (2011); Schwerdt and Woessmann (2017); Woessmann (2018)).

[8] For analyses of the effects of NCLB and predecessor reforms, see Hanushek and Raymond (2005), Jacob (2005), Neal and Schanzenbach (2010), Rockoff and Turner (2010), Dee and Jacob (2011), Rouse, Hannaway, Goldhaber, and Figlio (2013), Reback, Rockoff, and Schwartz (2014), and Deming, Cohodes, Jennings, and Jencks (2016); see Figlio and Loeb (2011) for a survey.

not for the students. However, even in these cases, strategic complementarity or substitutability in the effort levels of the different actors might produce some ambiguity in responses.[9]

Between the two ends of the continuum are standardized forms of testing that do not include external comparisons. For example, teachers may regularly use assessments in their classroom that are standardized rather than self-developed but that do not provide for a comparison to students in other schools or to the district or national average. In addition, use of standardized tests may support a variety of report card systems without external comparison. It is less obvious that this type of information would solve the described principal-agent problems.[10]

In general, our analysis of testing abstracts from the particular use to which the generated achievement information is put. However, there is one category of internal testing – measures aimed at teacher monitoring – that cannot be separated from a particular use. For example, consider inspections of teacher lessons set up to be used for the monitoring of teacher practices. We cannot identify whether it is the availability of testing per se or its particular use that is having an impact. Therefore, we will separate internal teacher monitoring out from other forms of internal testing in our empirical application below with the acknowledgment that this is not purely a category of information provision.

These considerations lead us to focus on four categories of testing: (1) standardized testing with external comparison, (2) standardized testing without external comparison, (3) internal testing, and (4) internal teacher monitoring.

While the discussion so far did not differentiate specific school environments, the policy uses of information from student testing across countries are unlikely to be uniform across

[9] For a general discussion, see Todd and Wolpin (2003) and De Fraja, Oliveira, and Zanchi (2010). Reback (2008) finds that students do respond in cases where their performance is important to school ratings.

[10] In prior work on the US, accountability that had consequential impacts on schools was more closely related to student performance than accountability confined to report card information (Hanushek and Raymond (2005)).

systems with different levels of institutional development.[11] For example, a set of high-performing schools might be expected to know how to react to achievement signals and different rewards. Therefore, they may react more strongly to any type of incentive structure created from student assessments than an otherwise comparable set of low-performing schools. But the results might also just be the opposite: Low-performing schools have more room for improvement and may be in greater need to have their incentives focused on student outcomes. High-performing schools, by contrast, may have the capacities and be subject to overall political and schooling institutions that already better reflect the desires of parents.

3. International Panel Data

To extract evidence on how test-based information affects student learning, we combine international measures of student achievement with measures of different types of student assessments over a period of 15 years. We describe each of the two components in turn.

3.1 Six Waves of PISA Student Achievement Tests

In 2000, the Organisation for Economic Co-operation and Development (OECD) conducted the first wave of the international achievement test called Programme for International Student Assessment (PISA). Since then, PISA has tested the math, science, and reading achievement of representative samples of 15-year-old students in all OECD countries and an increasing number of non-OECD countries on a three-year cycle (OECD (2016)).[12] PISA makes a concerted effort

[11] Another dimension of heterogeneity may be across parents within a system, in that parents differ in their value functions, discount rates, and/or capacity to drive favorable results. Such differences may lie behind movements such as parents opting out of state-wide testing in the US, as some parents may feel that the measured output does not provide much information about the type of achievement they care about.

[12] The target population contains all 15-year-old students irrespective of the educational institution or grade that they attend. Most countries employ a two-stage sampling design, first drawing a random sample of schools in which 15-year-old students are enrolled (with sampling probabilities proportional to schools' number of 15-year-old students) and second randomly sampling 35 students of the 15-year-old students in each school.

to ensure random sampling of schools and students and to monitor testing conditions in participating countries. Data are not reported for countries that do not meet the standards.[13] PISA does not follow individual students over time, but the repeated testing of representative samples of students creates a panel structure of countries observed every three years.

In our analyses, we consider student outcomes in all countries that have participated in at least three of the six PISA waves between 2000 and 2015.[14] This yields a sample of 59 countries (35 OECD and 24 non-OECD countries, see Appendix Table A1) observed in 303 country-by-wave observations. We perform our analysis at the individual student level, encompassing a total sample of 2,187,415 students in reading and slightly less in math and science.

PISA uses a broad set of tasks of varying difficulty to create a comprehensive indicator of the continuum of students' competencies in each of the three subjects. PISA assessments last for up to two hours. Using item response theory, achievement in each domain is mapped on a scale with a mean of 500 test-score points and a standard deviation of 100 test-score points for OECD-country students in the 2000 wave. The test scales are then psychometrically linked over time.[15] Until 2012, PISA employed paper and pencil tests. In 2015, the testing mode was changed to computer-based testing, a topic we will come back to in our robustness analysis below.

While average achievement across all countries was quite stable between 2000 and 2015, achievement has moved significantly up in some countries and significantly down in others (see Appendix Figure A1). In 14 countries, achievement improved by at least 20 percent of a standard

[13] In particular, due to deviations from the protocol, the data exclude the Netherlands in 2000, the United Kingdom in 2003, the United States in the reading test 2006, and Argentina, Kazakhstan, and Malaysia in 2015.

[14] We include the tests conducted in 2002 and 2010 in which several previously non-participating countries administered the 2000 and 2009 tests, respectively. We exclude any country-by-wave observation for which the entire data of a background questionnaire is missing. This applies to France from 2003-2009 (missing school questionnaire) and Albania in 2015 (missing student questionnaire). Liechtenstein was dropped due to its small size.

[15] The math (science) test was re-scaled in 2003 (2006), any effect of which should be captured by the year fixed effects included in our analysis.

deviation compared to their initial achievement (in decreasing order, Peru, Qatar, Brazil, Luxembourg, Chile, Portugal, Israel, Poland, Italy, Mexico, Indonesia, Colombia, Latvia, and Germany). On the other hand, achievement decreased by at least 20 percent of a standard deviation in eleven countries (United States, Korea, Slovak Republic, Japan, France, Netherlands, Finland, Iceland, United Kingdom, Australia, and New Zealand).

In student and school background questionnaires, PISA provides a rich array of background information on the participating students and schools. Students are asked to provide information on their personal characteristics and family background, and school principals provide information on the schools' resources and institutional setting. We select a set of core variables of student characteristics, family backgrounds, and school environments that are available in each of the six waves and merge them with the test score data into one dataset comprising all PISA waves. Student-level controls include student gender, age, first- and second-generation immigration status, language spoken at home, parental education (measured in six categories), parental occupation (four categories), and books at home (four categories). School-level controls include school size (number of students), community location (five categories), share of fully certified teachers, principals' assessments of the extent to which learning in their school is hindered by teacher absenteeism (four categories), shortage of math teachers, private operation, and share of government funding. At the country level, we include GDP per capita and, considering the results in Hanushek, Link, and Woessmann (2013), the share of schools with academic-content autonomy and its interaction with initial GDP per capita. We impute missing values in the student and school background variables by using the respective country-by-wave mean and include a set of indicators for each imputed variable-by-observation.[16]

[16] The share of missing values is generally very low for the covariates, see Appendix Table A2.

3.2 Categories of Testing

We derive our measures of different forms of student testing, consistently measured across countries and time, from a combination of the PISA school background questionnaires, regular data collection of other parts of the OECD, and data compiled under the auspices of the European Commission. This provides us with 13 separate indicators of testing practices, each measured at the country-by-wave level over the period 2000-2015.[17] We collapse this range of testing aspects into the four categories derived in our conceptual framework. Here we summarize the constructed categories; details of questions and sources are provided in the Data Appendix.

Standardized Testing with External Comparison. The first category draws on four separate data sources that identify standardized assessments constructed outside of schools and designed explicitly to allow comparisons of student outcomes across schools and students. This category includes the proportion of schools where (according to the principals of schools participating in PISA) performance of 15-year-olds is regularly compared through external examinations to students across the district or the nation (which we term "school-based external comparison"). It also includes indicators of whether central examinations affect student placement at the lower secondary level (two sources) and whether central exit exams determine student outcomes at the end of secondary school (which, together, we term "student-based external comparison").[18]

Standardized Testing without External Comparison. The second testing category refers to standardized assessments that do not necessarily provide for or are not primarily motivated by external comparison. Three questions in the PISA survey document the prevalence of different

[17] Appendix Table A3 provides an overview of the different underlying assessment indicators. Appendix Table A4 indicates the number of country observations by wave for each indicator.

[18] As discussed in the Data Appendix, data on assessments for student placement are available for only a subset of (largely OECD) countries.

aspects of this type of testing: standardized testing in the tested grade, student tests to monitor teacher practices, and tracking of achievement data by an administrative authority.

Internal Testing. This category covers testing used for general pedagogical management including informing parents of student progress, public posting of outcomes, and tracking school outcomes across cohorts. The included measures are derived from three separate PISA questions.

Internal Teacher Monitoring. This final category covers internal assessments that are directly focused on teachers. It combines schools' use of assessments to judge teacher effectiveness and the monitoring of teacher practice by principals and by external inspectorates, again derived directly from the principal surveys in PISA.

Aggregation of Separate Indicators. The original 13 indicators of assessment practices were aggregated into the four main categories as the simple average of the observed indicators in each category.[19] Constructing the aggregate categories serves several purposes. In various instances, the survey items are measuring very similar concepts within the same content area, so that the aggregation acts to reduce measurement error in the individual questions and to limit multicollinearity at the country level (which is key in our identification strategy). For example, as discussed more fully in the appendix, the correlation between the two measures of national standardized exams used in lower secondary school is 0.59 in our pooled dataset (at the country-by-wave level) and 0.54 after taking out country and year fixed effects (which reflects the identifying variation in our model). Similarly, the two internal-testing measures of using

[19] The variables in each category are calculated as proportionate usage in terms of the specific indicators for each country and wave. Note also that indicator data entirely missing for specific PISA waves are imputed by country-specific linear interpolation of assessment usages, a procedure that retains the entire country-by-wave information but that does not influence the estimated impact of the test category because of the inclusion of imputation dummies in the panel estimates (see Data Appendix for details). The fact that imputation is not affecting our results is also shown by their robustness to using only the original (non-imputed) observations for each of the underlying 13 separate indicators (see Table 4).

assessments to inform parents and to monitor school progress are correlated at 0.42 in the pooled data and 0.57 after taking out country and year fixed effects (all highly significant). Additionally, the aggregation permits including the added information from some more specialized OECD and EU sources while not forcing elimination of other countries outside these boundaries.[20]

Descriptive Statistics. Table 1 provides descriptive statistics for the individual indicators of student testing and for the four combined testing categories. The measures derived from the PISA background questionnaires are shares bounded between 0 and 1, whereas the other testing measures are dummy variables.[21] As is evident, some testing practices are more common than others. For example, 89 percent of schools in our country-by-wave observations use some form of assessment to inform parents, but only 29 percent have national standardized exams in lower secondary school. Appendix Table A1 provides country-by-country statistics of the initial and final value of the four separate indicators of standardized testing with external comparison.

For our estimation, the variation over time within individual countries in the different types of testing is key. Figure 1 shows histograms of the 15-year change in the combined measures of the four testing categories for the 38 countries observed in both the first and last PISA waves. The implicit policy changes across student assessments in the sampled countries are clearly substantial and supportive of our estimation strategy based on a country-level panel approach.[22]

[20] Note that a number of indicators draw on principals' responses about the use of tests in their own schools. Because the PISA sampling involves different schools in each wave, some random error could be introduced. The aggregation also helps to eliminate this sort of measurement error.

[21] In federal countries, the dummy variables capture whether the majority of the student population in a country is subject to the respective assessment policy.

[22] The exception in this depiction is internal testing. However, the reduction in this aggregate measure is fully accounted for by a change in the wording of the questionnaire item on assessments to inform parents, where the word "assessments" was replaced by the word "standardized tests" in the 2015 questionnaire (see Appendix Table A3). While the mean of this item hardly changed (from 0.98 to 0.97) between 2000 and 2012, it dropped to 0.64 in 2015. Ignoring the 2015 value, the mean of the combined measure of internal testing increased by 0.08 from 2000 to 2012. This example indicates the importance of including year fixed effects in our analyses and of taking particular care in considering the question wording. As we will show below, our qualitative results on internal testing are unaffected by dropping the year 2015 from the analysis.

Importantly, there is also wide variation in the change in the prevalence of the different forms of student assessments across countries, providing the kind of variation used for identification in our analysis. The policy variation is larger for standardized testing with external comparison than for the other three categories, leading us to expect higher precision (lower standard errors) of the coefficient estimates for this category.

The increasing use of external assessments is quite evident.[23] For example, the share of schools that are externally compared with student assessments increased by more than 50 percentage points in five countries (Luxembourg, Denmark, Italy, Portugal, and Poland) and by more than 20 percentage points in another 18 countries. In three countries, by contrast, the share decreased by more than 20 percentage points (Tunisia, Costa Rica, and Croatia).

No data source provides consistent external documentation of the time pattern of different legislated testing policies across countries. But we can rely upon the actual school implementation pattern identified by the principals at the time of each testing wave. Our interest is how different test-based information relates to student outcomes and does not seek to evaluate specific accountability or incentive policies that may be concurrently or subsequently introduced. Some changes in testing regimes have been directly related to more comprehensive reforms such as the 2006 *Folkeskole* Act in Denmark that introduced a stronger focus on assessment including national tests (Shewbridge, Jang, Matthews, and Santiago (2011)) and the introduction of standardized national assessments to monitor student outcomes in Luxembourg (Shewbridge, Ehren, Santiago, and Tamassia (2012)). But it appears more common that testing programs are introduced independent of any prescribed overall incentive or accountability system such as the

[23] Appendix Figure A2 depicts the evolution of using standardized assessments for school-based external comparison from 2000 to 2015 for each country.

2009 introduction of the *Invalsi* national test in Italy.[24] This information then plays into a variety of local uses by schools and parents.

As these measures are derived from survey responses by principals, they reflect the combined effect of external policies and the actual implementation of them at the school level. Thus, for example, the introduction of national assessments in Denmark is not accompanied by a discontinuous jump but by a more gradual implementation path.

4. Empirical Model

Identifying the impacts of testing in a cross-country analysis is of course challenging. Assessments are not exogenously distributed across schools and countries. At the student level, an obvious potential source of bias stems from the selection of otherwise high-performing students into schools that have specific assessment practices. At the country level, there may also be reverse causality if poorly performing countries introduce assessment systems in order to improve their students' achievement. Ultimately, any omitted variable that is associated both with the existence of student assessments and with student achievement levels will lead to bias in conventional estimation. In the cross-country setting, for example, unobserved country-level factors such as culture, the general valuation of educational achievement, or other government institutions may introduce bias.

We address leading concerns of bias in cross-country estimation by formulating a fixed-effects panel model of the following form:

$$A_{ict} = I_{ict}\alpha_I + S_{ict}\alpha_S + C_{ct}\alpha_C + T_{ct}\beta + \mu_c + \mu_t + \varepsilon_{ict} \quad (1)$$

[24] See Appendix Figure A2 and the description in https://it.wikipedia.org/wiki/Test_INVALSI.

Achievement A of student i in country c at time t is expressed as a linearly additive function of vectors of input factors at the level of students I, schools S, and countries C, as well as the measures of student testing T. The parameters μ_c and μ_t are country and year fixed effects, respectively, and ε_{ict} is an individual-level error term. We start by estimating separate models for each testing category and subsequently report models that consider all four categories simultaneously.

Our fixed-effects panel model identifies the effect of assessment practices on student achievement only from country-level within-country variation over time. First, note that the treatment variable, T_{ct}, is aggregated to the country-by-wave level. This specification avoids bias from within-country selection of students into schools that use student assessments. Second, we include country fixed effects, μ_c, to address any potential bias that arises from unobserved time-invariant country characteristics that may be correlated with both assessments and achievement. The specification exploits the fact that different countries have reformed their assessment systems at different points in time. Our parameters of interest β will not be affected by systematic, time-invariant differences across countries.[25] This specification implies that countries that do not change their assessment practices over the observation period will not enter into the estimation of β. The model also includes time fixed effects μ_t. These capture any global trends in achievement along with common shocks that affect testing in a specific PISA wave (including any changes in the testing instruments).

[25] Some recent investigations of scores on international assessments have focused on differential effort levels of students across countries (see, for example, Borghans and Schils (2012); Zamarro, Hitt, and Mendez (2016); Gneezy et al. (2017); Balart, Oosterveen, and Webbink (2018)). These differences in noncognitive effects related to our outcome variable of PISA scores would be captured by the country fixed effects as long as they do not interact with the incentives introduced by various applications of testing. Note also that other analysis that experimentally investigated test motivation effects in a short form of the very PISA test employed here did not find significant effects of informational feedback, grading, or performance-contingent financial rewards on intended effort, actual effort, or test performance (Baumert and Demmrich (2001)).

We think of this specification as a reduced-form model characterizing the impact of different kinds of performance information on the overall level of learning, *A*. Information per se does not change student outcomes unless it triggers different behavior from parents, students, and teachers. Any altered behavior could be the result of specific incentive programs or it could reflect an array of local and family responses to the information. Our purpose, however, is not to trace these different potential mechanisms but to understand the role of different kinds of assessment information. Sometimes test information is explicitly linked to specific incentives (such as the case of student exit exams), but more generally this is not the case.

The key identifying assumption of our model is the standard assumption of fixed-effects panel models. Conditional on the rich set of control variables at the student, school, and country level included in our model, in the absence of reform the change in student achievement in countries that have introduced or extended assessment practices would have been similar to the change in student achievement in countries that did not reform at the given point in time. In Section 6, we provide specification tests of this identifying assumption.

5. Results

This section presents our baseline results, as well as heterogeneous results by school environment. All models are estimated as panel models with country and year fixed effects, conditioning on the rich set of control variables at the student, school, and country level indicated above.[26] Regressions are weighted by students' sampling probabilities within countries, giving equal weight to each country-by-wave cell across countries and waves. Standard errors are clustered at the country level throughout.

[26] Appendix Table A1 shows the coefficients on all control variables for the specification of the first column in Table 5. The estimates for control variables are quite consistent across specifications.

5.1 Baseline Results on Internal and External Testing

Table 2 presents the results for the combined measures of the four testing categories, first entered separately (columns 1-4) and then jointly (columns 5-7). The basic impact results suggest that different forms of student testing have very different effects on student achievement. Among the four assessment categories, only changes in standardized testing that is used for external comparisons have a strong and statistically significant positive relationship with changes in student outcomes. The coefficients on standardized testing without external comparison and internal testing are insignificant and close to zero, whereas there is quite a sizeable negative coefficient on internal teacher monitoring.[27] These different impacts are consistent with the predictions on differing strengths of potential incentives from the conceptual discussion.

The point estimate for standardized testing with external comparison suggests that a change from not used to complete standardized external comparison is related to an increase in math achievement by more than one quarter of a standard deviation. The point estimates and the statistical significance of the category impacts are very similar between the regressions that include each testing category individually and the regression that includes all four categories simultaneously (column 5), indicating that there is enough independent variation in the different testing categories for estimation and that the effect of standardized external comparison does not reflect reforms in other assessment categories. In the inclusive regression, the negative coefficient on internal teacher monitoring even turns significant in math. With that nuanced

[27] Note that, consistent with the larger within-country variation of standardized testing with external comparison over time documented in section 3.2, the standard error associated with this coefficient estimate is smaller. Still, even with the smaller standard error of this variable, the coefficient estimates on standardized testing without external comparison and internal testing would be far from statistical significance.

exception, results for science and reading achievement are very similar to those for math (columns 6 and 7).[28]

Conceptually, the category of external comparisons actually aggregates two quite distinct components related to schools and students, respectively. One component considers standardized assessments for external comparison of schools to district or national performance. This category mainly indicates information created to spotlight school performance and potentially having its greatest effect on administrators and teachers. The second category combines three measures of testing to determine school and career placement decisions for students with the clear focus on the students themselves.

Table 3 disaggregates standardized testing with external comparison into school-based and student-based external comparisons.[29] The impact of both school and student assessments is strongly positive and statistically significant, with estimates for the school-based testing being somewhat larger than for the individual student testing. The results suggest that focusing information on different actors encourages different responses and leads to separate effects on outcomes. This table presents simultaneous estimates for the other three categories, none of which is qualitatively affected.

To establish that our aggregation is not suppressing important heterogeneity within the separate categories, Table 4 presents individual results for each of the 13 underlying country-

[28] The hypothesis that the effect of standardized testing with external comparison is the same as the effects of the other three testing categories is jointly strongly rejected in each of the three subjects. Individually, the coefficient on standardized testing with external comparison is significantly different from standardized testing without external comparison in math and reading, from internal testing in reading, and from internal teacher monitoring in all three subjects.

[29] The measure of student-based external comparison is the simple average of the three underlying indicators of standardized testing with external comparison except for the one on school-based external comparison. Note that the estimates of Table 3 are based on smaller student samples from fewer countries, because data on student-based external comparison are available for few countries beyond OECD and European Union countries.

level indicators of student assessment, where each cell represents a separate regression.[30] Of particular interest, each of the four elements of the external comparison composite, with one exception, has a significantly positive impact on student performance in the three subjects. The exception is the use of central exit examinations, which could simply reflect that student performance measured by PISA at age 15 is not very responsive to testing that only occurs at the end of secondary school (when students are usually aged around 18 or 19). While the point estimates are positive in all three subjects, they do not reach statistical significance.[31] The estimated coefficients for the other three indicators taken separately are smaller than the combined measure. As noted, this probably reflects a reduction in measurement error for the correlated indicators and the fact that the incentives created by the different assessments are not perfect substitutes, implying that the combined impact across components is greater than that for any individual component.[32]

At the individual indicator level in Table 4, there is also some evidence of positive effects of standardized testing in the relevant grade for PISA and some indication of impact from the use of assessment to inform parents in science. None of the other indicators of standardized testing without external comparison, of internal testing, and of internal teacher monitoring is significantly related to student achievement on average. Ignoring statistical significance, the

[30] The separate regressions of Table 4 do not employ any imputation of the separate treatment variables. Thus, the number of countries and waves included in each estimation varies and is determined by the availability of the specific testing indicator. The fact that these results confirm the previous results of the four combined categories shows that the latter are not driven by the interpolated imputations required for the aggregation of the separate indicators.

[31] Consistent with the weaker evidence on central exit exams, constructing the combined measure of standardized testing with external comparison without the central exit exam measure (i.e., based on the other three underlying indicators) yields a slightly larger coefficient estimate of 30.9 in the specification of column 5 of Table 2.

[32] A third possibility is that the estimation samples for the separate indicators are varied and smaller than for the combined indicator. However, we reject this explanation because estimating the combined model in column 5 of Table 2 just for the smallest sample of countries in the separate indicator models yields a virtually identical coefficient for standardized testing with external comparison.

point estimates suggest that the potential negative impact of the internal monitoring of teachers is driven by the two subjective components – monitoring by the school principal and by external inspectorates. The aggregate categorical variable is larger than these two subcomponents, potentially again reflecting a reduction in measurement error and possible additivity.

5.2 Environmental Differences in Informational Impact

Countries enter our observation period at very different stages of educational development, and almost certainly with environments that have both different amounts of information about schools and different degrees of policy interactions among parents, administrators, and teachers. One straightforward way to parameterize these differences is to explore how incentive effects vary with a country's initial level of achievement.

We introduce interaction terms between the testing measures T_{ct} and a country's average achievement level when it first participated in PISA, $\bar{A}_{c0}$:

$$A_{ict} = I_{ict}\alpha_I + S_{ict}\alpha_S + C_{ct}\alpha_C + T_{ct}\beta_1 + (T_{ct} \times \bar{A}_{c0})\beta_2 + \mu_c + \mu_t + \varepsilon_{ict} \qquad (2)$$

The parameters β_2 indicate whether the testing effect varies between countries with initially low or high performance. Note that the initial performance level is a country feature that does not vary over time, so that any main effect is captured by the country fixed effects μ_c included in the model.

Table 5 presents estimates of the interacted model for the three subjects. The left three columns provide results for the aggregate category of standardized testing with external comparison, while the right three columns divide the external comparisons into school-based and student-based comparisons. The initial score is centered on 400 PISA points (one standard

deviation below the OECD mean). The precise patterns of estimated effects by initial achievement with confidence intervals are displayed in Figure 2 for math performance.

The picture of how the overall achievement environment interacts with the impact of different forms of testing can be summarized as follows. First, the impact of standardized testing with external comparison is stronger in lower achieving countries and goes to zero for the highest achieving countries. In particular, at an initial country level of 400 PISA points the introduction of standardized external comparison leads to an increase in student achievement of 37.3 percent of a standard deviation in math. With each 100 initial PISA points, this effect is reduced by 24.6 percent of a standard deviation. At an initial level of 500 PISA points (the OECD mean), the effect of standardized external comparison is still statistically significantly positive at around 13 percent of a standard deviation in all three subjects. Second, standardized testing without external comparison similarly creates significant impact in initially low-achieving countries, with effects disappearing for higher-achieving countries (i.e., those with initial scores of roughly above 490 in all subjects). Third, the estimate of internal testing is insignificant throughout the initial-achievement support. Fourth, the estimates for internal teacher monitoring are insignificant for most of the initial-achievement distribution but turn negative only at high levels of initial achievement in math. Fifth, when external comparisons are disaggregated into school-based and student-based components, school-based comparisons follow essentially the same heterogeneous pattern as overall standardized testing with external comparison but go to zero for a somewhat larger set of initially high-achieving countries. By contrast, the impact of student-based external comparisons does not vary significantly with initial achievement levels.

The disaggregated underlying individual indicators of standardized testing with external comparison consistently show the pattern of significantly stronger effects in initially poorly

performing countries (Appendix Table A5).[33] Interestingly, the introduction of central exit exams – which did not show a significant effect on average – also shows the pattern of decreasing effects with higher initial achievement, in particular in science. Similarly, all three underlying indicators of standardized testing without external comparison also show the same pattern of significant positive effects at low levels of achievement and significantly decreasing effects with initial achievement. Thus, the positive effect of standardized testing in low-achieving countries appears to be quite independent of whether the standardized tests allow for external comparison or just for monitoring. This finding supports the World Bank attention to testing for low achieving countries (World Bank (2018)).[34]

In contrast to the significant interactions with initial achievement levels, we do not find evidence of consistent heterogeneities in several other environmental dimensions (not shown). In particular, the effects of the four testing categories do not significantly interact with countries' initial level of GDP per capita, which contrasts with the heterogeneous effects found for school autonomy in that dimension in Hanushek, Link, and Woessmann (2013). Similarly, there are no significant interactions of the testing categories with the level of school autonomy in a country. In addition, standardized testing with external comparison does not significantly interact with the other three categories of student assessments.

[33] There is no significant heterogeneity in the effect of the Eurydice measure of national testing, which is likely due to the fact that this measure is available only for 18 European countries which do not feature a similarly wide range of initial achievement levels.

[34] An interesting outlier in the individual-indicator analysis are assessments to inform parents, which show the opposite type of heterogeneity (significantly so in math and science): The expansion of assessments to inform parents about their child's progress does not have a significant effect at low levels of initial achievement, but the effect gets significantly more positive at higher levels. Among initially high-performing countries, informing parents leads to significant increases in student achievement. E.g., at an initial achievement level of 550 PISA points, there is a significantly positive effect on science achievement of 37.0 percent of a standard deviation. It seems that addressing assessments at parents is only effective in raising student achievement in environments that already show a high level of achievement, capacity, and responsiveness of schools.

6. Specification Tests

Our fixed-effects panel model identifies the effect of assessment policies on student achievement from policy changes within countries over time. In this section, we return to a discussion of the identifying assumptions of our specification and a series of tests of their validity.

6.1 A Placebo Test with Leads of the Testing Variables

A leading remaining concern of the fixed-effects model is that reforms may be endogenous, in the sense that reforming countries may already be on a different trajectory than non-reforming countries for other reasons, thus violating the usual common-trend assumption of the fixed-effects model. Here the largest concern is that countries that are on a downward trend turn to expanded testing to reform the system. Note that, if generally true, this would tend to bias our estimated effects downward.

Our panel setup lends itself to an informative placebo test. In particular, any given reform should *not* have a causal effect on the achievement of students in the wave *before* it is implemented. Including leads of the assessment measures – i.e., additional variables that indicate the assessment status in the *next* PISA wave – provides a placebo test of this.

As is evident in Table 6, none of the lead variables of the four testing categories is significantly related to student achievement (i.e., in the wave before reform implementation).[35] At the same time, the results of the contemporaneous testing measures are fully robust to conditioning on the lead variables: Standardized testing with external comparison has a significant positive effect on the math, science, and reading achievement of students *in the year*

[35] The coefficients on the lead variables are somewhat imprecisely estimated. However, in models with leads for just standardized testing with external comparison, the lead coefficient is statistically significantly different from the base coefficient at the 5 percent level in science, at the 10 percent level in reading, and at the 20 percent level in math.

in which it is implemented, but not in the wave in which it is not yet implemented. Moreover, the estimated coefficients for the testing categories are qualitatively similar to those in Table 2.[36]

The fact that the leads of the testing variables are insignificant also indicates that lagged achievement does not predict assessment reforms. In that sense, the results speak against the possibility that endogeneity of assessment reforms to how a school system is performing is a relevant concern for the interpretation of our results.

Estimating the full interacted model with all four testing categories and their leads interacted with initial achievement is overly demanding to the data. Nevertheless, focusing just on the main results of Section 5.2, an interacted model that includes just standardized testing with external comparison, its lead, and their interactions with initial achievement gives confirmatory results: standardized testing with external comparison is significantly positive, its interaction with initial achievement is significantly negative, and both the lead variable and its interaction with initial achievement are statistically insignificant (not shown).

No similar test is possible for the lag of the testing variables, as lagged testing policies may in fact partly capture the effect of previously implemented reforms to the extent that reforms take time to generate their full effects. In a specification that includes the contemporaneous, lead, and lagged variable, both the contemporaneous and the lag of the standardized testing with external comparison variable are statistically significant while the lead remains insignificant (not shown).

There is no evidence of the introduction of different testing regimes in response to prior educational circumstances. At the same time, it is clearly difficult to estimate time patterns reliably given that we are limited to at most six time-series observations for each country. Thus,

[36] By construction, the placebo regression with leads excludes the 2015 PISA data, so the most direct comparison would be the baseline model without the 2015 wave. As indicated in Table 9 below, results are very similar in that specification.

while highly suggestive, definitive testing of the key identifying assumptions such as common trends across countries is not possible.[37]

6.2 Coincidental Other Policies, Long Differences, and other Specification Tests

Another important possible remaining concern is that countries may introduce other policies coincidentally with the use of alternative testing policies. Although we cannot consider all such potential policy changes, we can directly analyze what is the most likely synchronized policy – expanded local autonomy in school decision making. Local schools have greater knowledge both of the demands they face and of their own capacities, making them attractive places for much decision making. But for just the reasons discussed in the conceptual model, with asymmetric information about their actions and results, they might not operate in an optimal way from the viewpoint of either the higher-level policy makers or even of the parents.

All our estimations include information on the time pattern of autonomy reforms for each country. Consistent with prior work (Hanushek, Link, and Woessmann (2013)), our results confirm that the effect of school autonomy on student achievement is negative in developing countries but positive in developed countries in this extended setting.[38] Importantly, the results on assessment effects are not confounded by the potentially coincidental introduction of policies that alter school decision making and autonomy.

As a further indication against the potential concern that other contemporaneous correlated policy changes might affect our results, note that results do not change when the four different testing categories are entered individually or jointly. That is, other forms of testing – and their

[37] For example, adding a linear time trend for each country renders coefficients too imprecise for clear inference.

[38] With six rather than four PISA waves and with 303 rather than 155 country-by-wave observations, we show here that the previous results about autonomy are also robust to the consideration of the effects of student assessment reforms (see Appendix Table A2).

potentially coinciding other policy changes – are controlled for in the simultaneous model. Only other policies that are coincidental just with the specific form of testing and not with the other ones could potentially still introduce bias. Furthermore, all models control for several time-varying school features including the schools' share of government funding, private/public management, and size. The school-level covariates also include several variables related to teachers – the share of fully certified teachers, teacher absenteeism, and shortage of math teachers. Contemporaneous policy reforms in these school features are thus also controlled for.

In fact, some of these school-level variables – in particular, those capturing the composition of teachers – could potentially be endogenous to the testing reforms. However, Table 7 shows that qualitative results in math are unaffected by leaving the teacher controls out of the specification (column 1). The same is true for achievement in science and reading (not shown).

Another approach to gauge the potential relevance of unobserved factors to affect our results is to look at the extent to which the inclusion of the entire set of observed factors changes our estimates. Dropping all covariates from the model does not change the qualitative results (column 2). This invariance holds despite the fact that the explained variance of the model increases substantially by the inclusion of the control variables, from 0.256 to 0.391. The fact that results are insensitive to the included set of relevant covariates reduces concerns that our estimates are strongly affected by any omitted variable bias from unobserved characteristics (in the sense of Altonji, Elder, and Taber (2005)).

Our fixed-effects panel model is identified from changes that occur from one PISA wave to the next, i.e., from three-year changes. This strategy has the advantage of incorporating several changes per country. The disadvantages are that any measurement error is amplified in the first-differenced changes and that any impact of testing may take time to emerge fully (as suggested

by the model with testing lags alluded to above). By restricting identification to changes across all sample periods, we can both reduce the potential influence of measurement error and gauge the long-run relevance of the policy reforms.

Column 3 of Table 7 provides estimates from a model in long differences that considers just the total 15-year change from the first to the last PISA wave. Our main findings are robust in this long-difference specification. Consistent with larger measurement error in shorter-frequency change data, the estimate of the positive effect of standardized testing with external comparison is larger when considering only long-run changes. The estimates of effects of the other three testing categories remain insignificant.

The long-difference analysis provides a convenient way to illustrate the main results about how changes in standardized external comparison translate into achievement gains. Figure 3 displays the added-variable plot for the impact of changes in standardized testing with external comparison. It clearly shows that countries that expanded the use of standardized external comparison from 2000 to 2015 saw the achievement of their students improve.

Relatedly, there is a difference between legislated testing reforms and the actual implementation of testing in schools. The latter is particularly relevant for understanding the impacts of actual testing usage, whereas the former may carry particular interest from a policy perspective. As discussed in section 3.2, the implementation path of test usage may be more gradual than any formal policy reform at the national level. Most of our testing measures are derived from reports of school principals on the use of testing in their schools, measured as the country share of schools using the specific testing application. But some are also dummy measures based on dichotomous coding of whether a country has formally legislated a specific testing policy or not, representing partial but well-measured policy changes. In particular, the

separate OECD and Eurydice measures of national standardized testing represent coding by country specialists of the changes in assessment policies – i.e., the kinds of accurately observed policy changes that would enter into micro policy evaluations.

While we prefer the combined testing measures in our baseline specification, it is important to note that the two dummy measures of standardized testing with external comparison are separately significant in their impact on overall student performance (see second and third lines in Table 4). Thus, the more gradual measure of usage of external comparison in schools and the discontinuous reform indicators of formal national policies yield very similar results, indicating that our results do not depend on adopting one of the specific perspectives.

As indicated in Table 8, also the results of the interacted specification are unaffected by dropping the teacher controls or all controls (columns 1 and 2). Similarly, while obviously less precise, the pattern of heterogeneity by initial achievement is also evident in the long-difference specification when the analysis is restricted to the category of standardized testing with external comparison (column 4).[39]

To check that the negative effects of standardized testing without external comparison and internal teacher monitoring at high levels of initial achievement (indicated in Figure 2) are not simply an artefact of the imposed linearity of the interaction model, columns 5-8 of Table 8 report results of a specification that interacts each of the four testing categories with four dummies reflecting the four quartiles of initial country achievement. There is no indication of strong nonlinearity.[40] In particular, the negative effects at high levels of initial achievement are

[39] Similarly, a model restricted to the category of standardized testing without external comparison yields a significantly positive main effect and a significantly positive interaction (not shown).

[40] The pattern for internal teacher monitoring also has a rather steady pattern when entered without the other three testing categories (92.3, -3.7, -36.6, and -102.5), suggesting that the joint specification with four interactions of four testing measures may be rather demanding to depict precise patterns. The separately estimated patterns for the other three measures also indicate rather linear relationships (not shown).

also visible in this specification, indicating that they are not driven by the imposition of linearity. This result may suggest that introducing standardized testing without external comparison and internal teacher monitoring in systems that are already performing at a high level may in fact detract teacher attention from more productive forms of instruction.

7. Robustness Analyses

Our results prove robust to a number of potentially contaminating factors. In particular, we consider possible peculiarities of our country sample, possible effects of student and school exclusions from PISA testing, possible interactions with changes in PISA testing, and an alternative two-stage estimation procedure. For ease of exposition, we present robustness results without heterogeneity by country achievement level in Table 9 in the text and the heterogeneity results, which yield similar conclusions, in Appendix Table A6.

To ensure that our results are not driven by the peculiarity of any specific country, we re-estimated all of our main models (the simultaneous regressions of columns 5-7 in Table 2 and columns 1-3 in Table 5) excluding one country at a time. The qualitative results are insensitive to this, with all significant coefficients remaining significant in all regressions (not shown).

To test whether results differ by level of development, we split the sample into OECD and non-OECD countries. As the first two columns of Table 9 show, qualitative results are similar in the two subgroups of countries, although the positive effect of standardized testing with external comparison is larger in OECD countries. Patterns of heterogeneity by achievement level are less precisely identified within the two more homogeneous subgroups (Appendix Table A6). In the OECD countries, the significant effect of standardized testing with external comparison does not vary significantly with initial achievement, but the demands of the fully interacted model make estimation difficult with just the 35-country sample. When we drop the insignificant interactions

(column 2), the point estimate of standardized testing with external comparison is significant. The heterogeneous effect of standardized testing without external comparison is somewhat more pronounced in OECD countries. But overall, the patterns do not differ substantively between the two country groups.

While PISA has stringent sampling standards, there is some variation across countries and time in the extent to which specific schools and students are excluded from the target population. Main reasons for possible exclusions are inaccessibility in remote regions or very small size at the school level as well as intellectual disability or limited test-language proficiency at the student level (OECD (2016)). The average total exclusion rate is below 3 percent, but it varies from 0 percent to 9.7 percent across countries and waves. To test whether this variation affects our analysis, column 3 in Table 9 (and column 4 in Appendix Table A6) controls for the country-by-wave exclusion rates reported in each PISA wave. As is evident, results are hardly affected.

In 2015, PISA instituted a number of major changes in testing methodology (OECD (2016)). Most importantly, PISA changed its assessment mode from paper-based to computer-based testing. In addition, a number of changes in the scaling procedure were undertaken, including changing from a one-parameter Rasch model to a hybrid of a one- and two-parameter model and changing the treatment of non-reached testing items. We performed three robustness tests to check whether these changes in testing methodology affect our results.

First, the simplest test of whether our analysis is affected by the 2015 changes in testing methodology is to drop the 2015 wave from our regressions. As is evident from column 4 in Table 9 (and column 5 in Appendix Table A6), qualitative results do not change when estimating the model just on the PISA waves from 2000 to 2012, indicating that our results cannot be driven by the indicated changes in testing mode.

Second, to address the changes in the psychometric scaling procedure, PISA recalculated countries' mean scores in the three subjects for all PISA waves since 2006 using the new 2015 scaling approach. In the final column of Table 9, we run our model with these rescaled country mean scores instead of the original individual scores as the dependent variable for the PISA waves 2006 to 2015. Again, qualitative results do not change, indicating that the changes in scaling approach do not substantively affect our analysis.

Third, while no similar analysis is possible for the change in testing mode, we analyzed whether countries' change in PISA achievement from paper-based testing in 2012 to computer-based testing in 2015 is correlated with a series of indicators of the computer familiarity of students and schools in 2012 that we derive from the PISA school and student background questionnaires. As indicated by Appendix Table A7, indicators of computer savviness in 2012 do not predict the change in test scores between 2012 and 2015 across countries. In particular, the change in countries' test achievement is uncorrelated with several measures of schools' endowment with computer hardware, internet connectivity, and software, as well as with several measures of students' access to and use of computers, internet, and software at home. The only exception is that the share of schools' computers that are connected to the internet is in fact *negatively* correlated with a country's change in science achievement, speaking against an advantage of computer-savvy countries profiting from the change in testing mode.

Finally, while we estimate all models at the individual student level, the main treatment varies only at the country-by-wave level. An alternative way of estimating our model is thus a two-stage estimation. The first stage is a student-level estimation that regresses test scores on all control variables. After collapsing the residuals of this first-stage estimation to the country-by-wave level, the second stage is a standard panel model that regresses these collapsed residuals on

the testing variables, including country and wave fixed effects. Appendix Table A8 shows that this two-stage model yields quantitatively very similar results to our main model.[41]

8. Conclusions

The extent of student testing and its usage in school operations have become items of heated debate in many countries, both developed and developing. Some express the view that high-stakes tests – meaning assessments that enter into reward and incentive systems for some individuals – are inappropriate (Koretz (2017)). Others argue that increased use of testing is essential for the improvement of educational outcomes (World Bank (2018)) and, by extension, of economic outcomes (Hanushek and Woessmann (2015); Hanushek, Schwerdt, Wiederhold, and Woessmann (2015)).

Many of these discussions, however, fail to distinguish between alternative forms of testing. And, most applications of expanded student assessments used for accountability purposes have not been adequately evaluated, largely because they have been introduced in ways that make clear identification of impacts very difficult. Critically, the expansion of national testing programs has faced a fundamental analytical issue of the lack of suitable counterfactuals.

Our analysis turns to international comparisons to address the key questions of which forms of student testing appear to induce changes that promote higher achievement. The conceptual framework behind the empirical analysis is a principal-agent model that motivates focusing on the strength of potential policies built on the assessment information generated by different forms of testing. The empirical analysis employs the increasingly plentiful international student achievement data that now move toward providing identification of consequential implications

[41] The same qualitative results also emerge when collapsing the original test scores (without residualizing) to the country-by-wave level (not shown), consistent with the insensitivity of our student-level results to the inclusion of controls (see Table 7).

of national testing.[42] Specifically, the six waves of the PISA test between 2000 and 2015 permit country-level panel estimation that relies on within-country over-time analysis of country changes in testing practices. We combine data across 59 countries to estimate how varying testing situations and applications affect student outcomes.

Focusing on international comparisons has both advantages and costs. A variety of testing policies that are introduced at the national level cannot be adequately evaluated within individual countries, but moving to cross-country evaluations requires dealing with a range of other possible influences on student outcomes. Some issues of measurement error, imprecise wording of questionnaire responses, and other possible influences on student outcomes are clearly difficult to address with complete certainty. But the richness of the existing data permits a variety of specification and robustness tests designed to illuminate the potential severity of the most significant issues of coincidental policies or programs.

Our results indicate that assessment systems that use standardized tests to compare outcomes across schools and students lead to greater student outcomes. These assessment systems tend to support consequential incentive and accountability regimes and to produce higher student achievement than those that use standardized tests without external comparison. They also tend to produce greater achievement results than systems relying on localized or subjective information that cannot be readily compared across schools and classrooms, which have little or negative impacts on student achievement. Moreover, both external comparisons aimed at schools and at students result in greater student learning. The impact of general comparisons of

[42] Interestingly, even the international testing – conducted on a voluntary basis in a low-stakes situation – has come under attack for potentially harming the educational programs of countries. Recent analysis, however, rejects this potential problem (Ramirez, Schofer, and Meyer (2018)).

standardized testing at the school level appears somewhat stronger than testing used to sort students across educational opportunities and subsequent careers.

Most interestingly from an international perspective is the finding that assessment systems are more important for school systems that are performing poorly. It appears that systems that are showing strong results know more about how to boost student performance and are less in need of strong information and accountability systems. Overall, the results from international comparisons of performance suggest that school systems gain from measuring how their students and schools are doing and where they stand in a comparative way. Comparative testing appears to allow for better incentives for performance and for rewarding those who are contributing most to educational improvement efforts.

References

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber. 2005. "Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools." *Journal of Political Economy* 113, no. 1: 151-184.

Andrab, Tahir, Jishnu Das, and Asim Ijaz Khwaja. 2017. "Report cards: The impact of providing school and child test scores on educational markets." *American Economic Review* 107, no. 6: 1535-1563.

Andrews, Paul, and coauthors. 2014. "OECD and Pisa tests are damaging education worldwide." *The Guardian*: https://www.theguardian.com/education/2014/may/06/oecd-pisa-tests-damaging-education-academics (accessed June 20, 2018).

Balart, Pau, Matthijs Oosterveen, and Dinand Webbink. 2018. "Test scores, noncognitive skills and economic growth." *Economics of Education Review* 63: 134-153.

Baumert, Jürgen, and Anke Demmrich. 2001. "Test motivation in the assessment of student skills: The effects of incentives on motivation and performance." *European Journal of Psychology of Education* 16, no. 3: 441-462.

Bergbauer, Annika B., Eric A. Hanushek, and Ludger Woessmann. 2018. "Testing." NBER Working Paper 24836. Cambridge, MA: National Bureau of Economic Research.

Bishop, John H. 1997. "The effect of national standards and curriculum-based exams on achievement." *American Economic Review* 87, no. 2: 260-264.

Bishop, John H., and Ludger Woessmann. 2004. "Institutional effects in a simple model of educational production." *Education Economics* 12, no. 1: 17-38.

Borghans, Lex, and Trudie Schils. 2012. The leaning tower of Pisa: Decomposing achievement test scores into cognitive and noncognitive components. Mimeo.

Braga, Michela, Daniele Checchi, and Elena Meschi. 2013. "Educational policies in a long-run perspective." *Economic Policy* 28, no. 73: 45-100.

Burgess, Simon, Deborah Wilson, and Jack Worth. 2013. "A natural experiment in school accountability: The impact of school performance information on pupil progress." *Journal of Public Economics* 106: 57-67.

Card, David. 1999. "The causal effect of education on earnings." In *Handbook of Labor Economics, Vol. 3A*, edited by Orley Ashenfelter and David Card. Amsterdam: North-Holland: 1801-1863.

De Fraja, Gianni, Tania Oliveira, and Luisa Zanchi. 2010. "Must try harder: Evaluating the role of effort in educational attainment." *Review of Economics and Statistics* 92, no. 3: 577-597.

Dee, Thomas S., and Brian A. Jacob. 2011. "The impact of No Child Left Behind on student achievement." *Journal of Policy Analysis and Management* 30, no. 3: 418-446.

Deming, David J., Sarah Cohodes, Jennifer Jennings, and Christopher Jencks. 2016. "School accountability, postsecondary attainment, and earnings." *Review of Economics and Statistics* 98, no. 5: 848-862.

Eurydice. 2009. *National testing of pupils in Europe: Objectives, organisation and use of results*. Brussels: European Commission; Education, Audiovisual and Culture Executive Agency (EACEA), Eurydice.

Eurydice. 2017. Online platform, ec.europa.eu/eurydice. Brussels: Education Audiovisual & Culture Executive Agency (EACEA), Eurydice Unit.

Figlio, David, and Susanna Loeb. 2011. "School accountability." In *Handbook of the Economics of Education, Vol. 3*, edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann. Amsterdam: North Holland: 383-421.

Figlio, David N., and Joshua Winicki. 2005. "Food for thought: The effects of school accountability plans on school nutrition." *Journal of Public Economics* 89, no. 2-3: 381-394.

Gneezy, Uri, John A. List, Jeffrey A. Livingston, Sally Sadoff, Xiangdong Qin, and Yang Xu. 2017. "Measuring success in education: The role of effort on the test itself." NBER Working Paper No. 24004. Cambridge, MA: National Bureau of Economic Research.

Hanushek, Eric A., Susanne Link, and Ludger Woessmann. 2013. "Does school autonomy make sense everywhere? Panel estimates from PISA." *Journal of Development Economics* 104: 212-232.

Hanushek, Eric A., and Margaret E. Raymond. 2005. "Does school accountability lead to improved student performance?" *Journal of Policy Analysis and Management* 24, no. 2: 297-327.

Hanushek, Eric A., Guido Schwerdt, Simon Wiederhold, and Ludger Woessmann. 2015. "Returns to skills around the world: Evidence from PIAAC." *European Economic Review* 73: 103-130.

Hanushek, Eric A., and Ludger Woessmann. 2011. "The economics of international differences in educational achievement." In *Handbook of the Economics of Education, Vol. 3*, edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann. Amsterdam: North Holland: 89-200.

Hanushek, Eric A., and Ludger Woessmann. 2015. *The knowledge capital of nations: Education and the economics of growth*. Cambridge, MA: MIT Press.

Holmstrom, Bengt, and Paul Milgrom. 1991. "Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design." *Journal of Law, Economics and Organization* 7: 24-52.

Hout, Michael, and Stuart W. Elliott, eds. 2011. *Incentives and test-based accountability in education*. Washington, DC: National Academies Press.

Jacob, Brian A. 2005. "Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools." *Journal of Public Economics* 89, no. 5-6: 761-796.

Jacob, Brian A., and Steven D. Levitt. 2003. "Rotten apples: An investigation of the prevalence and predictors of teacher cheating." *Quarterly Journal of Economics* 118, no. 3: 843-877.

Jürges, Hendrik, Kerstin Schneider, and Felix Büchel. 2005. "The effect of central exit examinations on student achievement: Quasi-experimental evidence from TIMSS Germany." *Journal of the European Economic Association* 3, no. 5: 1134-1155.

Koning, Pierre, and Karen van der Wiel. 2012. "School responsiveness to quality rankings: An empirical analysis of secondary education in the Netherlands." *De Economist* 160, no. 4: 339-355.

Koretz, Daniel. 2017. *The testing charade: Pretending to make schools better*. Chicago: University of Chicago Press.

Laffont, Jean-Jacques, and David Martimort. 2002. *The theory of incentives: The principal-agent model*. Princeton, NJ: Princeton University Press.

Leschnig, Lisa, Guido Schwerdt, and Katarina Zigova. 2017. "Central school exams and adult skills: Evidence from PIAAC." Unpublished manuscript, University of Konstanz.

Luedemann, Elke. 2011. "Intended and unintended short-run effects of the introduction of central exit exams: Evidence from Germany." In Elke Luedemann, *Schooling and the formation of cognitive and non-cognitive outcomes*. ifo Beiträge zur Wirtschaftsforschung 39. Munich: ifo Institute.

Neal, Derek, and Diane Whitmore Schanzenbach. 2010. "Left behind by design: Proficiency counts and test-based accountability." *Review of Economics and Statistics* 92, no. 2: 263-283.

Nunes, Luis C., Ana Balcão Reis, and Carmo Seabra. 2015. "The publication of school rankings: A step toward increased accountability?" *Economics of Education Review* 49: 15-23.

OECD. 2015. *Education at a glance 2015: OECD indicators*. Paris: Organisation for Economic Co-operation and Development.

OECD. 2016. *PISA 2015 results (volume I): Excellence and equity in education*. Paris: Organisation for Economic Co-operation and Development.

Pritchett, Lant. 2015. "Creating education systems coherent for learning outcomes: Making the transition from schooling to learning." RISE Working Paper 15/005. Oxford: Research on Improving Systems of Education (RISE).

Ramirez, Francisco O., Evan Schofer, and John W. Meyer. 2018. "International tests, national assessments, and educational development (1970-2012)." *Comparative Education Review* 62, no. 3: 344-364.

Reback, Randall. 2008. "Teaching to the rating: School accountability and the distribution of student achievement." *Journal of Public Economics* 92, no. 5-6: 1394-1415.

Reback, Randall, Jonah Rockoff, and Heather L. Schwartz. 2014. "Under pressure: Job security, resource allocation, and productivity in schools under No Child Left Behind." *American Economic Journal: Economic Policy* 6, no. 3: 207-241.

Rockoff, Jonah, and Lesley J. Turner. 2010. "Short-run impacts of accountability on school quality." *American Economic Journal: Economic Policy* 2, no. 4: 119-147.

Rouse, Cecilia Elena, Jane Hannaway, Dan Goldhaber, and David Figlio. 2013. "Feeling the Florida heat? How low-performing schools respond to voucher and accountability pressure." *American Economic Journal: Economic Policy* 5, no. 2: 251-281.

Schwerdt, Guido, and Ludger Woessmann. 2017. "The information value of central school exams." *Economics of Education Review* 56: 65-79.

Shewbridge, Claire, Melanie Ehren, Paulo Santiago, and Claudia Tamassia. 2012. *OECD Reviews of Evaluation and Assessment in Education: Luxembourg*. Paris: OECD.

Shewbridge, Claire, Eunice Jang, Peter Matthews, and Paulo Santiago. 2011. *OECD Reviews of Evaluation and Assessment in Education: Denmark*. Paris: OECD.

Todd, Petra E., and Kenneth I. Wolpin. 2003. "On the specification and estimation of the production function for cognitive achievement." *Economic Journal* 113, no. 485: F3-33.

Woessmann, Ludger. 2003. "Schooling resources, educational institutions, and student performance: The international evidence." *Oxford Bulletin of Economics and Statistics* 65, no. 2: 117-170.

Woessmann, Ludger. 2016. "The importance of school systems: Evidence from international differences in student achievement." *Journal of Economic Perspectives* 30, no. 3: 3-32.

Woessmann, Ludger. 2018. "Central exit exams improve student outcomes." *IZA World of Labor* 2018: 419.

Woessmann, Ludger, Elke Luedemann, Gabriela Schuetz, and Martin R. West. 2009. *School accountability, autonomy, and choice around the world*. Cheltenham, UK: Edward Elgar.

World Bank. 2018. *World Development Report 2018: Learning to realize education's promise*. Washington, DC: World Bank.

Zamarro, Gema, Collin Hitt, and Ildefonso Mendez. 2016. "When students don't care: Reexamining international differences in achievement and non-cognitive skills." EDRE Working Paper No. 2016-18. University of Arkansas.

Data Appendix: Sources and Construction of Testing Measures

We derive a series of measures of different forms student testing over the period 2000-2015 from the PISA school background questionnaires and other sources. Information on testing is classified into four categories with varying strength of generated incentives: standardized testing with external comparison, standardized testing without external comparison, internal testing, and internal teacher monitoring. We aggregate each assessment measure to the country-by-wave level. Below, we also discuss how we combine the different indicators into an aggregate measure for each of the four testing categories. Details on the precise underlying survey questions and any changes in question wording over time are found in Appendix Table A3.

A.1 Standardized Testing with External Comparison

Drawing on four different sources, we combine four separate indicators of standardized testing designed to allow for external comparisons.

First, from the PISA school background questionnaires, we measure the share of schools in each participating country that is subject to assessments for external comparison. In particular, school principals respond to the question, "In your school, are assessments of 15-year-old students used to compare the school to district or national performance?" Figure A2 provides a depiction of the evolution of this measure from 2000 to 2015 for each country.

Second, in the 2015 version of its Education at a Glance (EAG) publication, the OECD (2015) published an indicator of the existence of national/central examinations at the lower secondary level together with the year that is was first established. The data were collected by experts and institutions working within the framework of the OECD Indicators of Education Systems (INES) program in a 2014 OECD-INES Survey on Evaluation and Assessment. National examinations are defined as "standardized student tests that have a formal consequence

for students, such as an impact on a student's eligibility to progress to a higher level of education or to complete an officially-recognized degree" (OECD (2015), p. 483). According to this measure, five of the 37 countries with available data have introduced national standardized exams in lower secondary school between 2000 and 2015.[43]

Third, following a very similar concept, the Eurydice unit of the Education, Audiovisual and Culture Executive Agency (EACEA) of the European Commission provides information on the year of first full implementation of national testing in a historical overview of national testing of students in Europe (Eurydice (2009); see also Braga, Checchi, and Meschi (2013)). In particular, they classify national tests for taking decisions about the school career of individual students, including tests for the award of certificates, promotion at the end of a school year, or streaming at the end of primary or lower secondary school. We extend their measure to the year 2015 mostly based on information provided in the Eurydice (2017) online platform. During our period of observation, eight of the 18 European countries introduced national tests for career decisions and two abolished them.

Fourth, Leschnig, Schwerdt, and Zigova (2017) compile a dataset of the existence of central exit examinations at the end of secondary school over time for the 31 countries participating in the Programme for the International Assessment of Adult Competencies (PIAAC). They define central exit exams as "a written test at the end of secondary school, administered by a central authority, providing centrally developed and curriculum based test questions and covering core subjects." Following Bishop (1997), they do not include commercially prepared tests or university entrance exams that do not have direct consequences for students passing them. Central exit exams "can be organized either on a national level or on a regional level and must be

[43] In federal countries, all system-level indicator measures are weighted by population shares in 2000.

mandatory for all or at least the majority of a cohort of upper secondary school." We extend their time period, which usually ends in 2012, to 2015. Five of the 30 countries in our sample introduced central exit exams over our 15-year period, whereas two countries abandoned them.

A.2 Standardized Testing without External Comparison

Beyond externally comparative testing, the PISA school background questionnaire also provides three additional measures of standardized testing that allow for different types of monitoring but do not readily provide for external comparison.

First, school principals answer the question, "Generally, in your school, how often are 15-year-old students assessed using standardized tests?" Answer categories start with "never" and then range from "1-2 times a year" ("yearly" in 2000) to more regular uses. We code a variable that represents the share of schools in a country that use standardized testing at all (i.e., at least once a year).

Second, school principals provide indicators on the following battery of items: "During the last year, have any of the following methods been used to monitor the practice of teachers at your school?" Apart from a number of non-test-based methods of teacher practice monitoring, one of the items included in the battery is "tests or assessments of student achievement." We use this to code the share of schools in a country that monitors teacher practice by assessments.

Third, school principals are asked, "In your school, are achievement data used in any of the following accountability procedures?" One consistently recorded item is whether "achievement data are tracked over time by an administrative authority," which allows us to construct a measure of the share of schools in a country for which an administrative authority tracks achievement data. The reference to over-time tracking by administrations indicates that the achievement data are standardized to be comparable over time.

A.3 Internal Testing

The PISA school background questionnaire also provides information on three testing policies where tests are not necessarily standardized and are mostly used for pedagogical management.

In particular, school principals report on the prevalence of assessments of 15-year-old students in their school for purposes other than external comparisons. Our first measure of internal testing captures whether assessments are used "to inform parents about their child's progress." The second measure covers the use of assessments "to monitor the school's progress from year to year." Each measure is coded as the share of schools in a country using the respective type of internal assessments.

The question on use of achievement data in accountability procedures referred to above also includes an item indicating that "achievement data are posted publicly (e.g. in the media)." Our third measure thus captures the share of schools in a country where achievement data are posted publicly. In the questionnaire item, the public posting is rather vaguely phrased and is likely to be understood by school principals to include such practices as posting the school mean of the grade point average of a graduating cohort, derived from teacher-defined grades rather than any standardized test, at the school's blackboard.

A.4 Internal Teacher Monitoring

Finally, the PISA school background questionnaire provides three additional measures of internal monitoring that are all focused on teachers.

First, again reporting on the prevalence of assessments of 15-year-old students in their school, school principals report whether assessments are used "to make judgements about teachers' effectiveness."

The battery of methods used to monitor teacher practices also includes two types of assessments based on observations of teacher practices by other persons rather than on student achievement tests. Our second measure in this area captures the share of schools where the practice of teachers is monitored through "principal or senior staff observations of lessons." Our third measure captures whether "observation of classes by inspectors or other persons external to the school" are used to monitor the practice of teachers.

A.5 Constructing Combined Measures for the Four Testing Categories

Many of the separate testing indicators are obviously correlated with each other, in particular within each of the four groups of testing categories. For example, the correlation between the EAG measure of national standardized exams in lower secondary school and the Eurydice measure of national tests for career decisions is 0.59 in our pooled dataset (at the country-by-wave level) and 0.54 after taking out country and year fixed effects (which reflects the identifying variation in our model). Similarly, the two internal-testing measures of assessments to inform parents and assessments to monitor school progress are correlated at 0.42 in the pooled data and 0.57 after taking out country and year fixed effects (all highly significant).

While these correlations are high, there is also substantial indicator-specific variation. These differences may reflect slight differences in the concepts underlying the different indicators and different measurement error in the different indicators, but also substantive differences in the measured assessment dimensions. In our main analysis, we combine the individual indicators into one measure for each of the four testing categories, but in additional analyses we report results for each indicator separately.

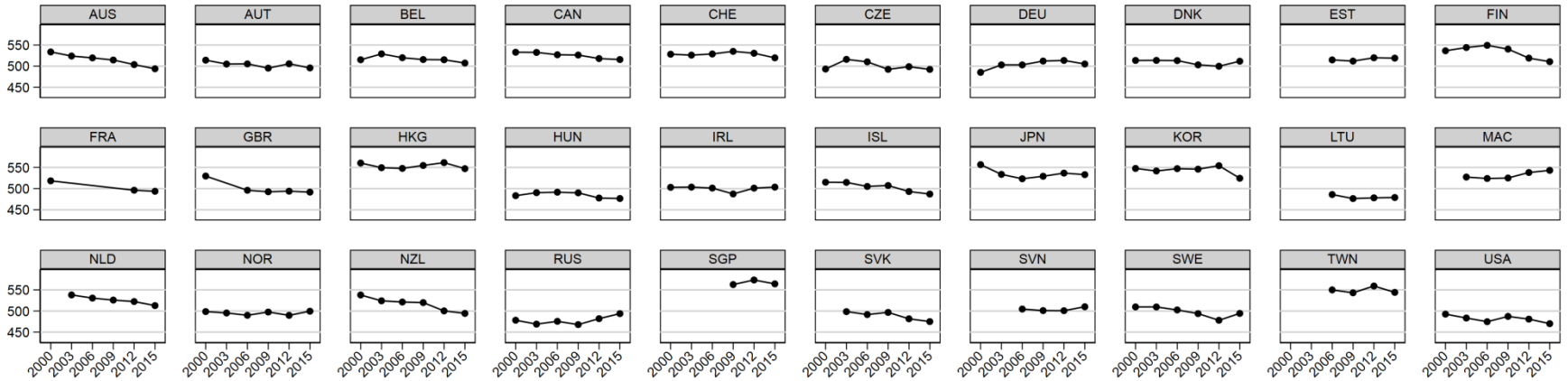
Our construction of the combined measures takes into account that the different indicators are available for different sets of waves and countries, as indicated in Appendix Table A4.

Before combining the indicators, we therefore impute missing observations in the aggregate country-by-wave dataset from a linear time prediction within each country. We then construct the combined measures of the four testing categories as the simple average of the individual imputed indicators in each category. To ensure that the imputation does not affect our results, all our regression analyses include a full set of imputation dummies that equal one for each underlying indicator that was imputed and zero otherwise.

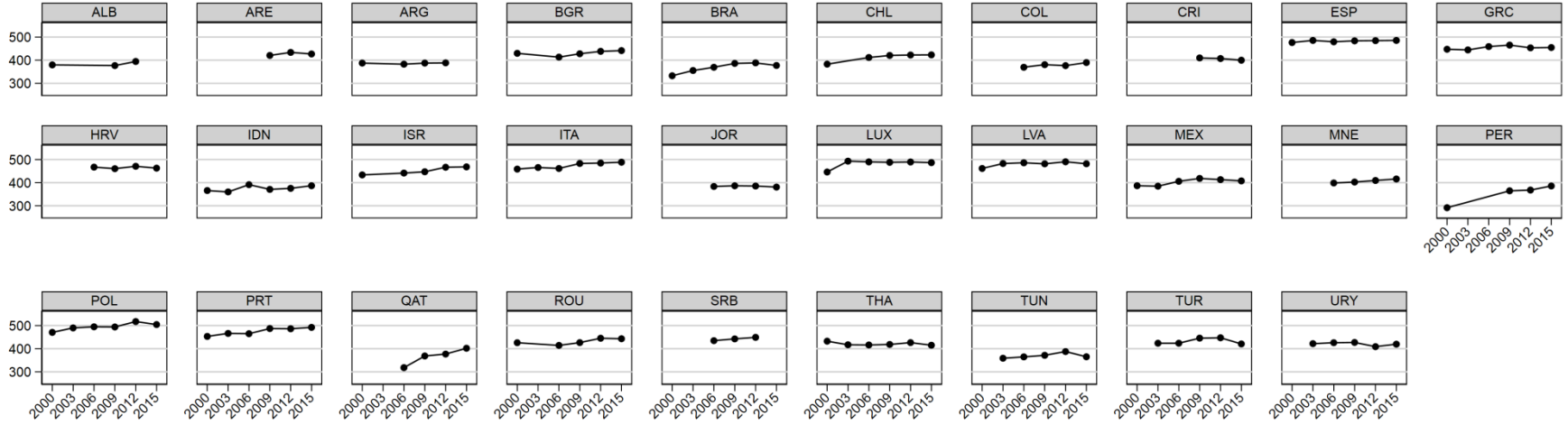
The combined measures of the four testing categories are also correlated with each other. In the pooled dataset of 303 country-by-wave observations, the correlations range from 0.278 between standardized testing with external comparison and internal teacher monitoring to 0.583 between standardized testing without external comparison and internal testing. After taking out country and year fixed effects, the correlations are lowest between standardized testing with external comparison and all other categories (all below 0.2), moderate between standardized testing without external comparison and the other categories (all below 0.3), and largest between internal testing and internal teacher monitoring (0.485). Because of potential multicollinearity, we first run our analyses for each aggregate assessment category separately and then report a model that considers all four categories simultaneously.

Figure A1: PISA math achievement in 2000-2015

Panel A: Countries above initial median achievement

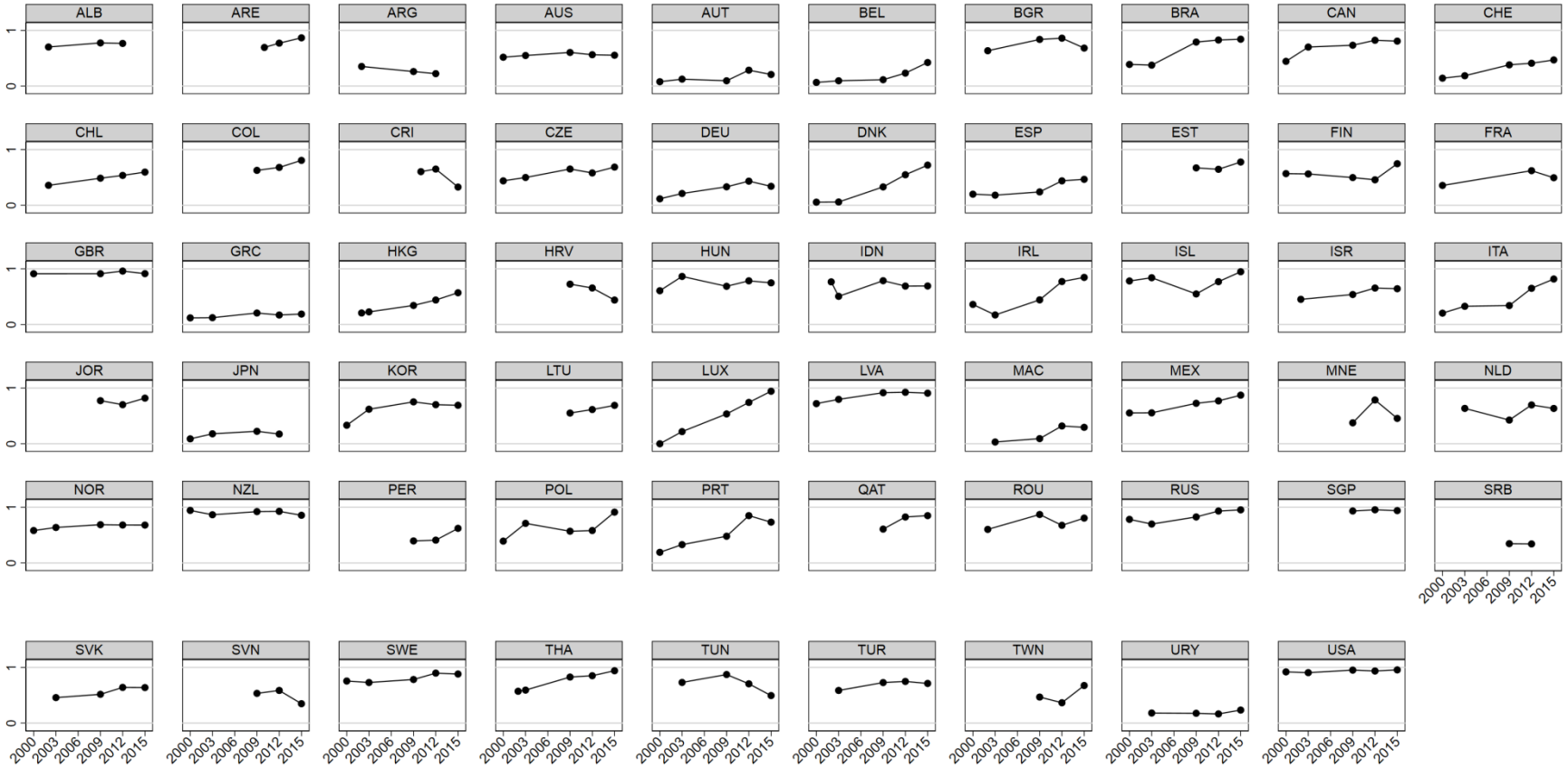


Panel B: Countries below initial median achievement



Notes: Country mean achievement in PISA math test. Country sample split at median of initial achievement level for expositional reasons. Country identifiers are listed in Appendix Table A1. Own depiction based on PISA micro data.

Figure A2: School-based external comparison in 2000-2015



Notes: Country share of schools with assessments for external comparison. Country identifiers are listed in Appendix Table A1. Own depiction based on PISA micro data.

Table A1: Selected indicators by country

	OECD	PISA math score		School-based external comparison		National standardized exams in lower sec. school		National tests for career decisions		Central exit exams	
	2015	2000	2015	2000	2015	2000	2015	2000	2015	2000	2015
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Albania (ALB) [a]	0	380	395	0.70	0.77	.	.	.	.	.	.
Argentina (ARG) [a]	0	387	389	0.35	0.22	.	.	.	.	.	.
Australia (AUS)	1	534	494	0.52	0.55	0	0	.	.	0.80	1
Austria (AUT)	1	514	496	0.08	0.21	0	0	.	.	0	0
Belgium (BEL)	1	515	507	0.07	0.42	0	0.32	0	0.32	.	.
Brazil (BRA)	0	333	377	0.39	0.84	0	0	.	.	.	.
Bulgaria (BGR) [a]	0	430	442	0.64	0.68	.	.	0	1	.	.
Canada (CAN)	1	533	516	0.44	0.81	0	0	.	.	0.54	0.54
Chile (CHL) [a]	1	383	423	0.36	0.60	0	0	.	.	0	0
Colombia (COL) [c]	0	370	390	0.63	0.81	0	0	.	.	.	.
Costa Rica (CRI) [e]	0	410	400	0.61	0.33	.	.	.	.	.	.
Croatia (HRV) [c]	0	467	463	0.73	0.44	.	.	.	.	.	.
Czech Republic (CZE)	1	493	492	0.44	0.69	0	0	0	0	0	1
Denmark (DNK)	1	514	512	0.06	0.72	1	1	1	1	1	1
Estonia (EST) [c]	1	515	519	0.67	0.78	1	1	.	.	1	0
Finland (FIN)	1	536	511	0.57	0.75	0	0	.	.	1	1
France (FRA)	1	518	494	0.36	0.50	1	1	.	.	1	1
Germany (DEU)	1	485	505	0.12	0.34	.	.	0	1	0.43	0.95
Greece (GRC)	1	447	455	0.12	0.19	0	0	0	0	1	0
Hong Kong (HKG) [a]	0	560	547	0.21	0.57	.	.	.	.	.	.
Hungary (HUN)	1	483	477	0.61	0.75	0	0	.	.	.	.
Iceland (ISL)	1	515	487	0.78	0.95	0	0	1	0	.	.
Indonesia (IDN) [a]	0	366	387	0.77	0.69	.	.	.	.	1	1
Ireland (IRL)	1	503	504	0.36	0.85	1	1	1	1	1	1
Israel (ISR) [a]	1	434	468	0.45	0.64	0	0	.	.	1	1
Italy (ITA)	1	459	489	0.21	0.82	1	1	0	1	1	1
Japan (JPN)	1	557	533	0.09	0.17	0	0	.	.	1	1
Jordan (JOR) [c]	0	384	381	0.77	0.82	.	.	.	.	.	.
Korea (KOR)	1	548	524	0.33	0.69	0	0	.	.	1	1
Latvia (LVA)	1	462	482	0.72	0.91	1	1	1	1	.	.

(continued on next page)

Table A1 (continued)

	OECD	PISA math score		School-based external comparison		National standardized exams in lower sec. school		National tests for career decisions		Central exit exams	
	2015	2000	2015	2000	2015	2000	2015	2000	2015	2000	2015
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Lithuania (LTU) [c]	0	486	479	0.55	0.69	.	.	0	0	1	1
Luxembourg (LUX) [b]	1	446	487	0.00	0.94	0	0	1	1	.	.
Macao (MAC)	0	527	543	0.03	0.30	.	.	.	.	.	.
Mexico (MEX)	1	387	408	0.55	0.87	0	0	.	.	.	.
Montenegro (MNE) [c]	0	399	416	0.38	0.46	.	.	.	.	.	.
Netherlands (NLD) [b]	1	538	513	0.64	0.63	1	1	1	1	1	1
New Zealand (NZL)	1	538	494	0.94	0.86	0	0	.	.	1	1
Norway(NOR)	1	499	500	0.58	0.68	0	1	0	1	1	1
Peru (PER) [a]	0	292	386	0.40	0.62	.	.	.	.	.	.
Poland (POL)	1	471	505	0.39	0.91	0	1	0	1	0	1
Portugal (PRT)	1	453	493	0.19	0.73	0	1	0	1	.	.
Qatar (QAT) [c]	0	318	402	0.61	0.85	.	.	.	.	.	.
Romania (ROU) [a]	0	426	443	0.60	0.81	.	.	0	1	.	.
Russia (RUS)	0	478	494	0.78	0.95	.	.	.	.	.	.
Serbia (SRB) [c]	0	435	449	0.35	0.34	.	.	.	.	.	.
Singapore (SGP) [d]	0	563	564	0.93	0.94	.	.	.	.	1	1
Slovak Republic (SVK) [b]	1	499	475	0.46	0.64	0	0	.	.	0	1
Slovenia (SVN) [c]	1	505	510	0.54	0.35	0	0	0	0	1	1
Spain (ESP)	1	476	486	0.20	0.47	0	0	.	.	0	0
Sweden (SWE)	1	510	494	0.76	0.88	0	0	1	1	0	0
Switzerland (CHE)	1	528	520	0.14	0.47	.	.	.	.	.	.
Taiwan (TWN) [c]	0	550	544	0.47	0.68	.	.	.	.	.	.
Thailand (THA) [a]	0	433	415	0.57	0.94	.	.	.	.	.	.
Tunisia (TUN) [b]	0	359	365	0.73	0.50	.	.	.	.	.	.
Turkey (TUR) [b]	1	424	421	0.59	0.71	1	1	.	.	0	0
United Arab Emirates (ARE) [e]	0	421	427	0.69	0.87	.	.	.	.	.	.
United Kingdom (GBR)	1	530	492	0.91	0.91	0	0	0.87	0	1	1
United States (USA)	1	493	470	0.92	0.96	0	1	.	.	0.07	0.07
Uruguay (URY) [b]	0	422	420	0.18	0.24	.	.	.	.		
Country average	0.59	465	469	0.48	0.66	0.23	0.35	0.39	0.67	0.66	0.72

Notes: PISA data: Country means, based on non-imputed data for each variable, weighted by sampling probabilities. "." = not available. [a-e] "2000" PISA data refer to country's initial PISA participation in [a] 2002, [b] 2003, [c] 2006, [d] 2009, [e] 2010.

Table A2: Descriptive statistics and complete model of basic interacted specification

	Descriptive statistics			Basic model	
	Mean	Std. dev.	Share imputed	Coeff.	Std. err.
Standardized testing with external comparison				37.304***	*(6.530)*
× initial score				-0.246***	*(0.085)*
Standardized testing without external comparison				67.772***	*(17.139)*
× initial score				-0.776***	*(0.175)*
Internal testing				-13.858	*(12.216)*
× initial score				0.161	*(0.100)*
Internal teacher monitoring				10.432	*(25.005)*
× initial score				-0.478*	*(0.249)*
Student and family characteristics					
Female	0.504	0.500	0.001	-11.557***	*(0.946)*
Age (years)	15.78	0.295	0.001	12.284***	*(0.921)*
Immigration background					
Native student	0.892				
First generation migrant	0.054	0.221	0.034	-8.322	*(4.635)*
Second generation migrant	0.054	0.223	0.034	-2.772	*(2.736)*
Other language than test language or national dialect spoken at home	0.111	0.305	0.061	-15.133***	*(2.309)*
Parents' education					
None	0.088	0.278	0.031		
Primary	0.019	0.134	0.031	9.138***	*(2.228)*
Lower secondary	0.062	0.238	0.031	10.814***	*(2.421)*
Upper secondary I	0.108	0.307	0.031	20.951***	*(2.984)*
Upper secondary II	0.077	0.262	0.031	26.363***	*(2.559)*
University	0.265	0.435	0.031	36.135***	*(2.538)*
Parents' occupation					
Blue collar low skilled	0.08	0.265	0.041		
Blue collar high skilled	0.088	0.278	0.041	8.401***	*(1.153)*
White collar low skilled	0.168	0.366	0.041	15.520***	*(1.108)*
White collar high skilled	0.335	0.464	0.041	35.601***	*(1.552)*
Books at home					
0-10 books	0.174	0.374	0.026		
11-100 books	0.478	0.493	0.026	30.297***	*(1.908)*
101-500 books	0.276	0.442	0.026	64.817***	*(2.426)*
More than 500 books	0.072	0.255	0.026	73.718***	*(3.433)*

(continued on next page)

Table A2 (continued)

	Descriptive statistics			Basic model	
	Mean	Std. dev.	Share imputed	Coeff.	Std. err.
School characteristics					
Number of students	849.0	696.7	0.093	0.012***	*(0.002)*
Privately operated	0.193	0.383	0.071	7.500*	*(4.396)*
Share of government funding	0.802	0.289	0.106	-16.293***	*(4.596)*
Share of fully certified teachers at school	0.822	0.294	0.274	6.662**	*(2.793)*
Shortage of math teachers	0.202	0.394	0.041	-5.488***	*(1.031)*
Teacher absenteeism					
No	0.337	0.427	0.213		
A little	0.484	0.447	0.213	-0.325	*(1.175)*
Some	0.140	0.310	0.213	-6.089***	*(1.556)*
A lot	0.039	0.173	0.213	-7.715***	*(2.413)*
School's community location					
Village or rural area (<3,000)	0.092	0.281	0.056		
Town (3,000-15,000)	0.208	0.397	0.056	5.238***	*(1.768)*
Large town (15,000-100,000)	0.311	0.451	0.056	9.935***	*(2.148)*
City (100,000-1,000,000)	0.251	0.422	0.056	14.209***	*(2.594)*
Large city (>1,000,000)	0.137	0.336	0.056	17.482***	*(3.447)*
Country characteristics					
Academic-content autonomy	0.597	0.248	-	-11.666	*(8.826)*
Academic-content autonomy × Initial GDP p.c.	5.043	7.578	-	1.871***	*(0.475)*
GDP per capita (1,000 $)	27.30	20.80	-	0.009	*(0.123)*
Country fixed effects; year fixed effects				Yes	
Student observations	2,193,026			2,094,856	
Country observations	59			59	
Country-by-wave observations	303			303	
R^2				0.393	

Notes: Descriptive statistics: Mean: international mean (weighted by sampling probabilities). Std. dev.: international standard deviation. Share imputed: share of missing values in the original data, imputed in the analysis. Basic model: Full results of the specification reported in first column of Table 5. Dependent variable: PISA math test score. Least squares regression weighted by students' sampling probability. Regression includes imputation dummies. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent.

Table A3: Measures of student testing: Sources and definitions

	Source (1)	Countries (2)	Waves (3)	Definition (4)	Deviation in wording in specific waves (5)
Standardized testing with external comparison					
School-based external comparison	PISA school questionnaire	PISA sample	2000-2003, 2009-2015	In your school, are assessments of 15-year-old students used for any of the following purposes? To compare the school to district or national performance.	2000: without "for any of the following purposes"; 2009-2015: "students in <national modal grade for 15-year-olds>" instead of "15-year-old students"; 2015: "standardized tests" instead of "assessments".
National standardized exams in lower secondary school	OECD (2015)	OECD EAG sample	2000-2015	National/central examinations (at the lower secondary level), which apply to nearly all students, are standardized tests of what students are expected to know or be able to do that have a formal consequence for students, such as an impact on a student's eligibility to progress to a higher level of education or to complete an officially recognized degree.	
National tests for career decisions	Eurydice (2009)	EU countries	2000-2015	Year of first full implementation of national testing, ISCED levels 1 and 2: Tests for taking decisions about the school career of individual pupils, including tests for the award of certificates, or for promotion at the end of a school year or streaming at the end of ISCED levels 1 or 2.	
Central exit exams	Leschnig, Schwerdt, and Zigova (2017)	PIAAC sample	2000-2015	Exit examination at the end of secondary school: A central exam is a written test at the end of secondary school, administered by a central authority, providing centrally developed and curriculum based test questions and covering core subjects. (See text for additional detail.)	
Standardized testing without external comparison					
Standardized testing in tested grade	PISA school questionnaire	PISA sample	2000, 2003, 2009, 2015	Generally, in your school, how often are 15-year-old students assessed using standardized tests? More than "never."	2009-2015: "students in <national modal grade for 15-year-olds>" instead of "15-year-old students"; 2009: "using the following methods:" "standardized tests"; 2015: "using the following methods:" "mandatory standardized tests" or "non-mandatory standardized tests".
Student tests to monitor teacher practice	PISA school questionnaire	PISA sample	2003, 2009-2015	During the last year, have any of the following methods been used to monitor the practice of teachers at your school? Tests or assessments of student achievement.	2003 and 2012: "mathematics teachers" instead of "teachers"; 2009: "<test language> teachers" instead of "teachers"
Achievement data tracked by administrative authority	PISA school questionnaire	PISA sample	2006-2015	In your school, are achievement data used in any of the following accountability procedures? Achievement data are tracked over time by an administrative authority.	

(continued on next page)

Table A3 (continued)

	Source (1)	Countries (2)	Waves (3)	Definition (4)	Deviation in wording in specific waves (5)
Internal testing					
Assessments to inform parents	PISA school questionnaire	PISA sample	2000-2003, 2009-2015	In your school, are assessments of 15-year-old students used for any of the following purposes? To inform parents about their child's progress.	2000: without "for any of the following purposes"; 2009-2015: "students in <national modal grade for 15-year-olds>" instead of "15-year-old students"; 2015: "standardized tests" instead of "assessments".
Assessments to monitor school progress	PISA school questionnaire	PISA sample	2000-2003, 2009-2015	In your school, are assessments of 15-year-old students used for any of the following purposes? To monitor the school's progress from year to year.	2000: without "for any of the following purposes"; 2009-2015: "students in <national modal grade for 15-year-olds>" instead of "15-year-old students"; 2015: "standardized tests" instead of "assessments".
Achievement data posted publicly	PISA school questionnaire	PISA sample	2006-2015	In your school, are achievement data used in any of the following accountability procedures? Achievement data are posted publicly (e.g. in the media).	
Internal teacher monitoring					
Teacher effective-ness judged by assessments	PISA school questionnaire	PISA sample	2000-2003, 2009-2015	In your school, are assessments of 15-year-old students used for any of the following purposes? To make judgements about teachers' effectiveness.	2000: without "for any of the following purposes"; 2009-2015: "students in <national modal grade for 15-year-olds>" instead of "15-year-old students"; 2015: "standardized tests" instead of "assessments".
Teacher practice monitored by principal	PISA school questionnaire	PISA sample	2003, 2009-2015	During the last year, have any of the following methods been used to monitor the practice of teachers at your school? Principal or senior staff observations of lessons.	2003 and 2012: "mathematics teachers" instead of "teachers"; 2009: "<test language> teachers" instead of "teachers"
Teacher practice monitored by external inspectors	PISA school questionnaire	PISA sample	2003, 2009-2015	During the last year, have any of the following methods been used to monitor the practice of teachers at your school? Observation of classes by inspectors or other persons external to the school.	2003 and 2012: "mathematics teachers" instead of "teachers"; 2009: "<test language> teachers" instead of "teachers"

Notes: Own depiction based on indicated sources.

Table A4: Country observations by wave

	2000/02 (1)	2003 (2)	2006 (3)	2009/10 (4)	2012 (5)	2015 (6)	Total (7)
Standardized testing with external comparison							
School-based external comparison	39	37	–	58	59	55	248
National standardized exams in lower secondary school	30	29	35	35	36	36	201
National tests for career decisions	17	15	21	21	21	21	116
Central exit exams	23	22	28	29	30	30	162
Standardized testing without external comparison							
Standardized testing in tested grade	38	35	–	58	–	51	182
Student tests to monitor teacher practice	–	36	–	57	59	56	208
Achievement data tracked by administrative authority	–	–	53	58	59	56	226
Internal testing							
Assessments to inform parents	40	37	–	58	59	55	249
Assessments to monitor school progress	40	37	–	58	59	55	249
Achievement data posted publicly	–	–	53	58	59	56	226
Internal teacher monitoring							
Teacher effectiveness judged by assessments	40	37	–	58	59	55	249
Teacher practice monitored by principal	–	37	–	58	59	56	210
Teacher practice monitored by external inspectors	–	37	–	58	59	56	210

Notes: Own depiction based on PISA data and other sources. See Data Appendix for details.

Table A5: Estimations for separate underlying testing indicators: Interacted specification

	Math		Science		Reading	
	Main effect	× initial score	Main effect	× initial score	Main effect	× initial score
	(1)	(2)	(3)	(4)	(5)	(6)
Standardized testing with external comparison						
School-based external comparison	39.945***	-0.456***	43.605***	-0.484***	47.018***	-0.481***
	(10.118)	*(0.078)*	*(10.441)*	*(0.117)*	*(9.023)*	*(0.098)*
National standardized exams in lower secondary school	50.625**	-0.464**	50.720***	-0.434**	39.186	-0.273
	(18.887)	*(0.206)*	*(13.905)*	*(0.162)*	*(31.246)*	*(0.301)*
National tests for career decisions	21.890***	-0.081	11.309	-0.002	20.983**	-0.119
	(5.524)	*(0.077)*	*(6.728)*	*(0.083)*	*(8.517)*	*(0.102)*
Central exit exams	24.550	-0.254	58.473***	-0.542***	54.899	-0.540
	(31.796)	*(0.322)*	*(18.255)*	*(0.156)*	*(46.933)*	*(0.543)*
Standardized testing without external comparison						
Standardized testing in tested grade	46.491***	-0.460***	42.679***	-0.427***	54.278***	-0.509***
	(9.608)	*(0.108)*	*(9.829)*	*(0.105)*	*(9.918)*	*(0.104)*
Student tests to monitor teacher practice	15.863	-0.384***	44.530***	-0.508***	25.154*	-0.391***
	(14.109)	*(0.116)*	*(14.908)*	*(0.174)*	*(12.715)*	*(0.130)*
Achievement data tracked by administrative authority	28.970*	-0.417***	38.054**	-0.419**	43.775**	-0.631**
	(14.631)	*(0.129)*	*(18.191)*	*(0.198)*	*(19.113)*	*(0.242)*
Internal testing						
Assessments to inform parents	-8.895	0.233***	-10.140	0.314***	-6.900	0.151
	(6.714)	*(0.047)*	*(8.012)*	*(0.079)*	*(10.352)*	*(0.103)*
Assessments to monitor school progress	6.106	-0.065	2.356	0.065	6.433	-0.115
	(8.812)	*(0.115)*	*(13.376)*	*(0.177)*	*(13.825)*	*(0.177)*
Achievement data posted publicly	15.898	-0.197	22.711	-0.264*	-8.159	-0.123
	(15.782)	*(0.133)*	*(15.355)*	*(0.144)*	*(19.472)*	*(0.236)*
Internal teacher monitoring						
Teacher effectiveness judged by assessments	0.387	-0.063	0.220	0.037	1.141	-0.043
	(14.989)	*(0.153)*	*(16.015)*	*(0.202)*	*(14.510)*	*(0.163)*
Teacher practice monitored by principal	0.807	-0.239	31.735	-0.514**	1.358	-0.186
	(26.483)	*(0.208)*	*(21.136)*	*(0.201)*	*(20.928)*	*(0.222)*
Teacher practice monitored by external inspectors	18.086	-0.370**	17.783	-0.365*	-6.485	-0.134
	(12.412)	*(0.145)*	*(17.744)*	*(0.207)*	*(16.606)*	*(0.189)*

Notes: Two neighboring cells present results of one separate regression, with "main effect" reporting the coefficient on the variable indicated in the left column and "× initial score" reporting the coefficient on its interaction with the country's PISA score in the initial year (centered at 400, so that the "main effect" coefficient shows the effect of assessments on test scores in a country with 400 PISA points in 2000). Dependent variable: PISA test score. Least squares regression weighted by students' sampling probability, including country and year fixed effects. Student assessment measures aggregated to the country level. Sample: student-level observations in six PISA waves 2000-2015. See Table 2 for the included control variables and Table 4 for numbers of observations, countries, and waves. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent.

Table A6: Robustness tests: Interacted specification

	OECD countries		Non-OECD countries	Control for exclusion rates	Without 2015	Rescaled test scale
	(1)	(2)	(3)	(4)	(5)	(6)
Standardized testing with external comparison	51.462	22.346***	26.378***	35.439***	35.085***	60.655***
	(30.820)	*(7.479)*	*(5.872)*	*(7.362)*	*(9.954)*	*(15.693)*
× initial score	-0.359		-0.374***	-0.217**	-0.189	-0.507**
	(0.326)		*(0.106)*	*(0.096)*	*(0.125)*	*(0.196)*
Standardized testing without external comparison	58.619*	64.291*	20.508	61.292***	55.777***	8.894
	(32.496)	*(34.495)*	*(18.675)*	*(20.757)*	*(19.008)*	*(30.447)*
× initial score	-0.547*	-0.636*	-0.319*	-0.716***	-0.703***	-0.152
	(0.321)	*(0.343)*	*(0.185)*	*(0.207)*	*(0.209)*	*(0.274)*
Internal testing	18.179	6.054	-10.840	-11.153	-1.941	-5.212
	(29.982)	*(11.613)*	*(13.040)*	*(12.372)*	*(31.980)*	*(15.369)*
× initial score	-0.134		0.232**	0.126	0.020	0.076
	(0.262)		*(0.105)*	*(0.105)*	*(0.334)*	*(0.131)*
Internal teacher monitoring	46.444	61.681	0.663	4.894	8.063	-72.152**
	(38.979)	*(40.538)*	*(20.416)*	*(29.938)*	*(40.220)*	*(35.725)*
× initial score	-0.733*	-0.887*	-0.342	-0.402	-0.681	0.666*
	(0.385)	*(0.387)*	*(0.315)*	*(0.292)*	*(0.434)*	*(0.359)*
Control variables	Yes	Yes	Yes	Yes	Yes	Yes
Country fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Student observations	1,434,355	1,434,355	660,501	2,045,454	1,679,250	1,698,971
Country observations	35	35	24	59	59	58
Country-by-wave observations	197	197	106	289	247	223
R^2	0.285	0.285	0.443	0.389	0.400	n.a.

Notes: Dependent variable: PISA math test score. Least squares regression weighted by students' sampling probability, including country and year fixed effects. Student assessment measures aggregated to the country level. Initial score: country's PISA score in the initial year (centered at 400, so that main-effect coefficient shows effect of assessments on test scores in a country with 400 PISA points in 2000). Sample: student-level observations in six PISA waves 2000-2015. Rescaled test scale available for waves 2006-2015 only. See Table 2 for included control variables. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent.

Table A7: Correlation of computer indicators in 2012 with change in PISA score from 2012 to 2015 at the country level

	Math (1)	Science (2)	Reading (3)
School			
Ratio of computers for education to students in respective grade	-0.015 *(0.912)*	-0.045 *(0.744)*	0.091 *(0.503)*
Share of computers connected to Internet	-0.223* *(0.099)*	-0.395*** *(0.003)*	-0.125 *(0.360)*
School's capacity to provide instruction hindered by:			
Shortage or inadequacy of computers for instruction	0.000 *(0.998)*	0.028 *(0.837)*	-0.029 *(0.834)*
Lack or inadequacy of Internet connectivity	0.106 *(0.438)*	0.247* *(0.066)*	0.040 *(0.771)*
Shortage or inadequacy of computer software for instruction	0.091 *(0.503)*	0.059 *(0.666)*	0.083 *(0.541)*
Student			
Computer at home for use for school work	0.034 *(0.805)*	0.240* *(0.075)*	-0.162 *(0.233)*
Number of computers at home	0.083 *(0.544)*	-0.043 *(0.751)*	0.181 *(0.182)*
Educational software at home	-0.111 *(0.414)*	0.044 *(0.746)*	-0.238* *(0.077)*
Link to the Internet at home	0.043 *(0.752)*	0.221 *(0.102)*	-0.116 *(0.394)*
Frequency of programming computers at school and outside of school	-0.150 *(0.270)*	-0.110 *(0.419)*	-0.003 *(0.980)*
Weekly time spent repeating and training content from school lessons by working on a computer	0.095 *(0.485)*	0.071 *(0.604)*	0.030 *(0.826)*

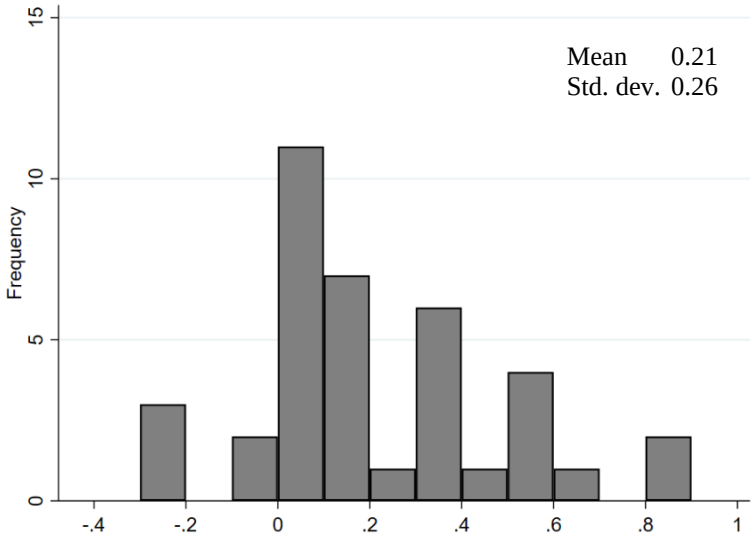
Notes: Correlation between the respective computer indicator (2012) indicated in the first column with the change in PISA test scores (2012-215) in the subject indicated in the header. Sample: 56 country-level observations of countries participating in the PISA waves 2012 and 2015. *p*-values in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent.

Table A8: Two-stage estimation: Panel model estimated at country-by-wave level

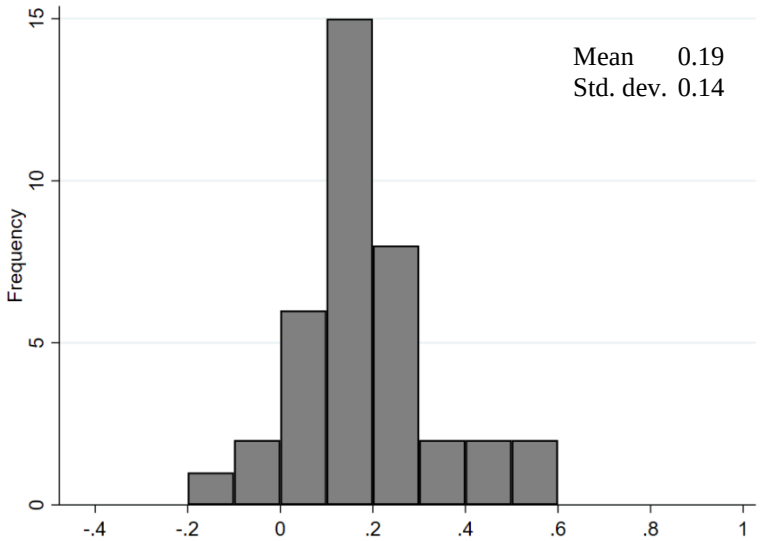
	Math	Science	Reading
	(1)	(2)	(3)
Standardized testing with external comparison	30.756***	24.357***	27.046***
	(7.236)	*(7.472)*	*(6.621)*
Standardized testing without external comparison	-4.765	0.402	-1.317
	(16.974)	*(17.391)*	*(14.641)*
Internal testing	5.404	15.201	-11.428
	(15.291)	*(17.128)*	*(17.067)*
Internal teacher monitoring	-36.953**	-31.555*	-26.154
	(18.188)	*(16.476)*	*(17.414)*
Country fixed effects	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes
Country observations	59	59	59
Country-by-wave observations	303	303	303

Notes: Dependent variable: country-level aggregation of the residuals of a first-stage student-level regression that regresses the PISA test score in the subject indicated in the header on student gender, age, parental occupation, parental education, books at home, immigration status, language spoken at home, school location, school size, share of fully certified teachers at school, teacher absenteeism, shortage of math teachers, private vs. public school management, share of government funding at school, country's GDP per capita, school autonomy, GDP-autonomy interaction, imputation dummies, country fixed effects and year fixed effects. Least squares regression at country-by-wave level, including country and year fixed effects. Sample: country-level observations in six PISA waves 2000-2015. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent.

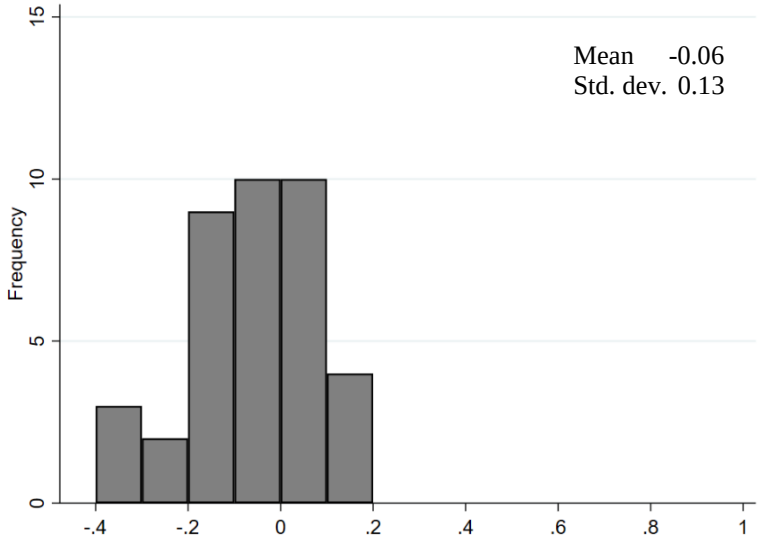
Figure 1: Histograms of change in four categories of student testing, 2000-2015



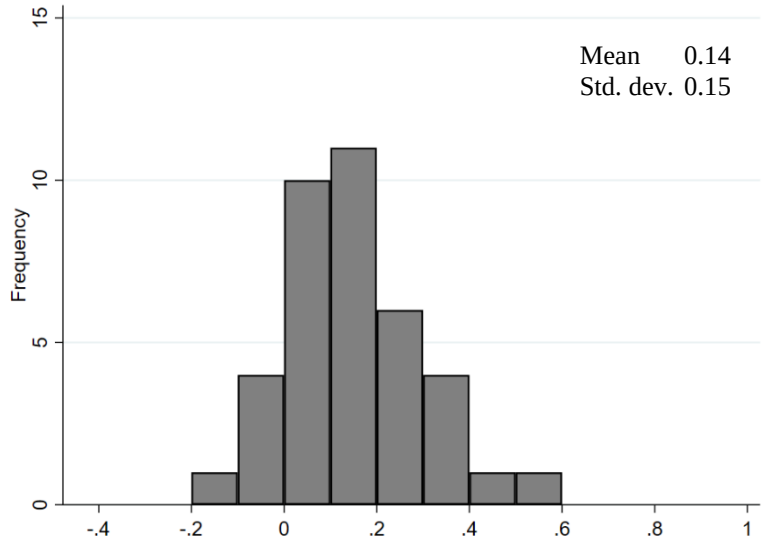
Standardized testing with external comparison



Standardized testing without external comparison



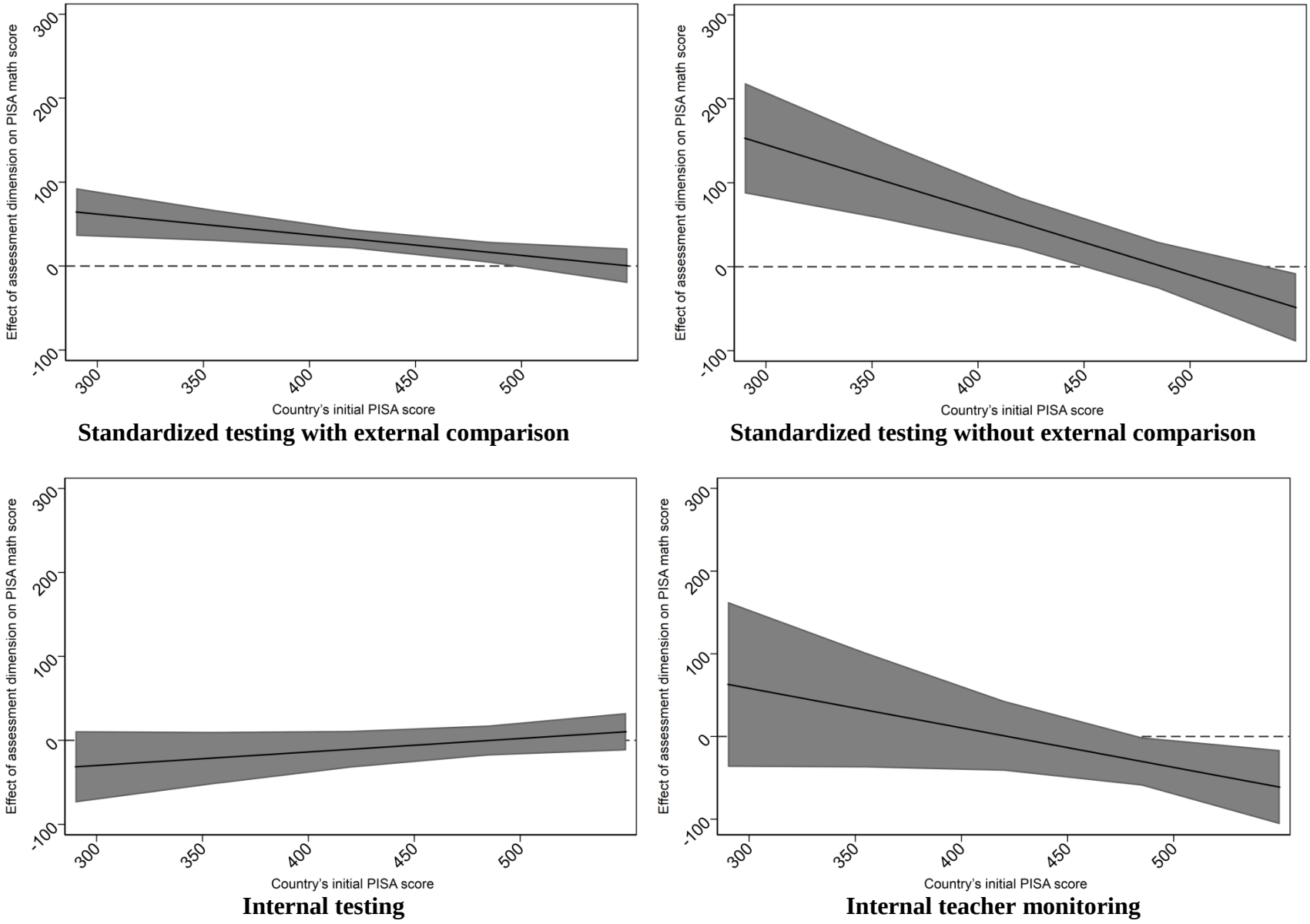
Internal testing



Internal teacher monitoring

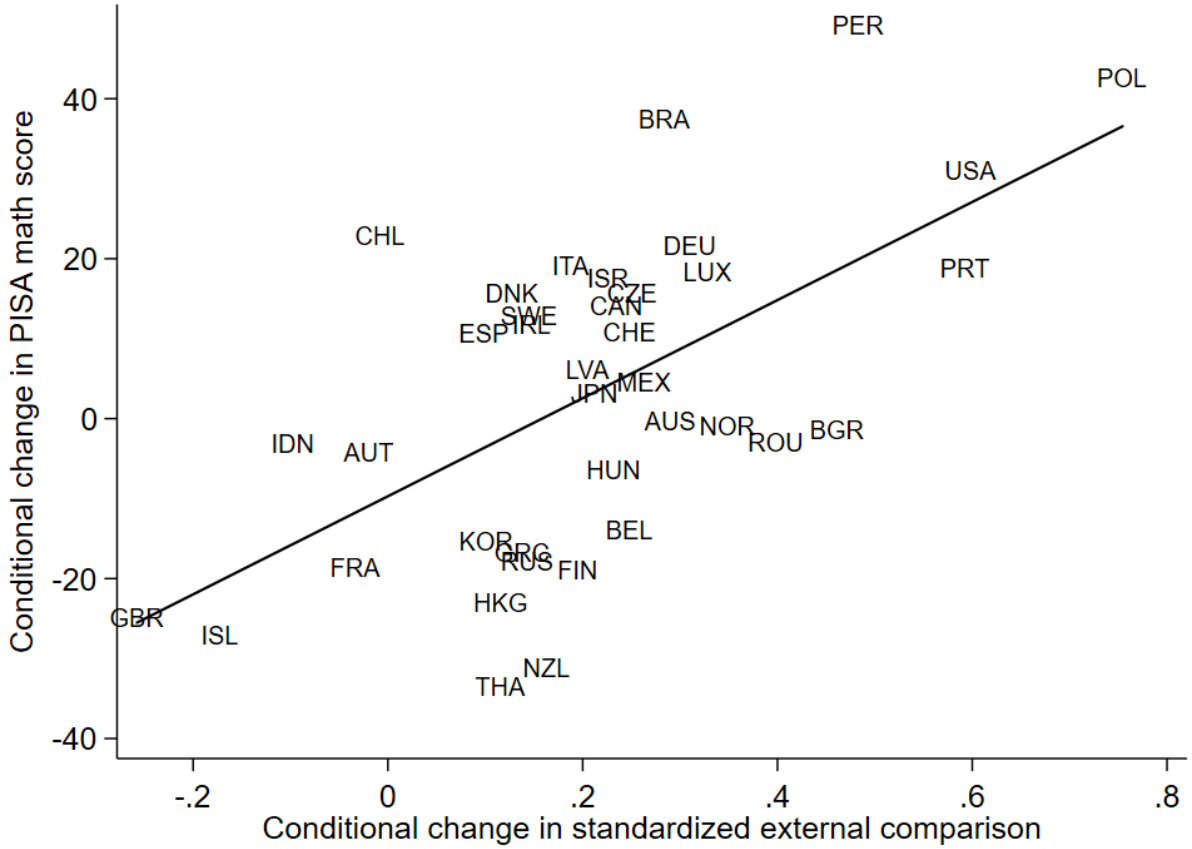
Notes: Histograms of change between 2000 and 2015 in the four combined measures of student assessment for the 38 countries observed both in the first and last PISA waves.

Figure 2: Effect of student testing on math performance by initial achievement levels



Notes: Average marginal effects of student assessments on PISA math score by initial country achievement, with 95 percent confidence intervals. See first column of Table 5 for underlying model.

Figure 3: Fifteen-year changes in standardized external comparison and in student achievement



Notes: Added-variable plot of the change in countries' average PISA math score between 2000 and 2015 against the change in the prevalence of standardized testing for external comparison, both conditional on a rich set of student, school, and country controls, based on a long-difference fixed-effect panel model estimated at the individual student level. Mean of unconditional change added to each axis. See column 3 of Table 7 for underlying model.

Table 1: Descriptive statistics of testing measures

	Mean (1)	Std. dev. (2)	Min (3)	Max (4)	Countries (5)	Waves (6)
Standardized testing with external comparison	**0.518**	**0.271**	**0.022**	**0.978**	**59**	**6**
School-based external comparison	0.573	0.251	0	0.960	59	5
National standardized exams in lower secondary school	0.292	0.452	0	1	37	6
National tests for career decisions	0.601	0.481	0	1	18	6
Central exit exams	0.689	0.442	0	1	30	6
Standardized testing without external comparison	**0.714**	**0.160**	**0.219**	**0.996**	**59**	**6**
Standardized testing in tested grade	0.721	0.233	0	1	59	4
Student tests to monitor teacher practice	0.750	0.191	0.128	1	59	4
Achievement data tracked by administrative authority	0.723	0.201	0.070	1	59	4
Internal testing	**0.684**	**0.147**	**0.216**	**0.963**	**59**	**6**
Assessments to inform parents	0.892	0.185	0.141	1	59	5
Assessments to monitor school progress	0.770	0.209	0	1	59	5
Achievement data posted publicly	0.393	0.239	0.016	0.927	59	4
Internal teacher monitoring	**0.553**	**0.216**	**0.026**	**0.971**	**59**	**6**
Teacher effectiveness judged by assessments	0.532	0.261	0	0.992	59	5
Teacher practice monitored by principal	0.773	0.262	0.049	1	59	4
Teacher practice monitored by external inspectors	0.402	0.255	0.006	0.994	59	4

Notes: Own depiction based on PISA micro data and other sources. See Data Appendix for details.

Table 2: The effect of different forms of student testing on student achievement: Fixed-effects panel models

	Math					Science	Reading
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Standardized testing with external comparison	26.365*** *(6.058)*				28.811*** *(6.126)*	23.282*** *(6.144)*	28.424*** *(5.911)*
Standardized testing without external comparison		-4.800 *(15.238)*			-5.469 *(14.062)*	1.252 *(13.950)*	-2.036 *(13.148)*
Internal testing			2.093 *(10.067)*		7.491 *(11.646)*	17.669 *(13.155)*	-12.660 *(14.736)*
Internal teacher monitoring				-23.478 *(14.518)*	-35.850** *(15.680)*	-27.549* *(14.226)*	-25.358 *(15.835)*
Control variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Country fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Student observations	2,094,856	2,094,856	2,094,856	2,094,856	2,094,856	2,094,705	2,187,415
Country observations	59	59	59	59	59	59	59
Country-by-wave observations	303	303	303	303	303	303	303
R^2	0.391	0.390	0.390	0.390	0.391	0.348	0.357

Notes: Dependent variable: PISA test score in subject indicated in the header. Least squares regression weighted by students' sampling probability, including country and year fixed effects. Student assessment measures aggregated to the country level. Sample: student-level observations in six PISA waves 2000-2015. Control variables include: student gender, age, parental occupation, parental education, books at home, immigration status, language spoken at home; school location, school size, share of fully certified teachers at school, teacher absenteeism, shortage of math teachers, private vs. public school management, share of government funding at school; country's GDP per capita, school autonomy, GDP-autonomy interaction; imputation dummies; country fixed effects; year fixed effects. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent.

Table 3: Disaggregation of standardized external comparison into school-based and student-based comparison

	Math	Science	Reading
	(1)	(2)	(3)
School-based external comparison	25.015***	21.317**	23.480***
	(7.667)	*(8.246)*	*(7.291)*
Student-based external comparison	17.309***	15.198***	14.481***
	(3.620)	*(3.883)*	*(3.753)*
Standardized testing without external comparison	-4.658	-8.333	-8.400
	(16.599)	*(15.007)*	*(14.602)*
Internal testing	4.896	13.419	-16.890
	(13.686)	*(15.306)*	*(18.616)*
Internal teacher monitoring	-35.424**	-27.374	-18.372
	(15.165)	*(16.656)*	*(16.373)*
Control variables	Yes	Yes	Yes
Country fixed effects	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes
Student observations	1,672,041	1,671,914	1,751,351
Country observations	42	42	42
Country-by-wave observations	230	230	230
R^2	0.348	0.315	0.321

Notes: Dependent variable: PISA test score in subject indicated in the header. Least squares regression weighted by students' sampling probability, including country and year fixed effects. Student assessment measures aggregated to the country level. Sample: student-level observations in six PISA waves 2000-2015. See Table 2 for included control variables. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent.

Table 4: Baseline model for separate underlying testing indicators

	Math (1)	Science (2)	Reading (3)	Observations (4)	Countries (5)	Waves (6)	R^2 (7)
Standardized testing with external comparison							
School-based external comparison	13.797*	13.147*	16.058**	1,703,142	59	5	0.382
	(7.417)	*(6.598)*	*(6.227)*				
National standardized exams in lower secondary school	13.400**	14.272**	14.568**	1,517,693	36	6	0.326
	(5.508)	*(5.336)*	*(5.418)*				
National tests for career decisions	15.650***	11.144***	11.002***	676,732	21	6	0.264
	(1.701)	*(2.377)*	*(2.932)*				
Central exit exams	3.694	8.242	9.806	1,141,162	30	6	0.308
	(7.041)	*(6.575)*	*(6.551)*				
Standardized testing without external comparison							
Standardized testing in tested grade	15.497**	11.051	19.380***	1,198,463	59	4	0.386
	(7.244)	*(6.901)*	*(7.169)*				
Student tests to monitor teacher practice	-19.266*	0.305	-10.046	1,537,802	59	4	0.385
	(9.625)	*(9.785)*	*(6.329)*				
Achievement data tracked by administrative authority	-3.555	5.173	-1.677	1,713,976	59	4	0.394
	(9.266)	*(9.578)*	*(12.787)*				
Internal testing							
Assessments to inform parents	7.923	14.664**	4.234	1,705,602	59	5	0.385
	(6.594)	*(6.974)*	*(7.912)*				
Assessments to monitor school progress	1.480	7.283	-1.598	1,705,602	59	5	0.385
	(5.343)	*(7.630)*	*(7.308)*				
Achievement data posted publicly	0.344	0.571	-16.954	1,713,976	59	4	0.394
	(8.371)	*(7.630)*	*(10.165)*				
Internal teacher monitoring							
Teacher effectiveness judged by assessments	-4.065	3.110	-1.981	1,705,602	59	5	0.385
	(8.249)	*(9.619)*	*(7.810)*				
Teacher practice monitored by principal	-19.751	-10.893	-14.239	1,588,962	59	4	0.385
	(14.072)	*(10.793)*	*(10.062)*				
Teacher practice monitored by external inspectors	-13.152	-13.524	-17.553*	1,588,962	59	4	0.385
	(10.038)	*(8.898)*	*(10.306)*				

Notes: Each cell presents results of a separate regression. Dependent variable: PISA test score. Least squares regression weighted by students' sampling probability, including country and year fixed effects. Student assessment measures aggregated to the country level. Sample: student-level observations in six PISA waves 2000-2015. See Table 2 for included control variables. Number of observations and R^2 refer to the math specification. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent.

Table 5: Effects of student testing by initial achievement level: Fixed-effects panel models

	Math (1)	Science (2)	Reading (3)	Math (4)	Science (5)	Reading (6)
Standardized testing with external comparison	37.304***	28.680***	47.977***			
	(6.530)	*(8.222)*	*(9.005)*			
× initial score	-0.246***	-0.149	-0.345***			
	(0.085)	*(0.101)*	*(0.113)*			
School-based external comparison				45.740***	39.343*	49.581**
				(15.067)	*(21.244)*	*(21.699)*
× initial score				-0.385**	-0.347	-0.361
				(0.165)	*(0.229)*	*(0.248)*
Student-based external comparison				15.138**	7.120	2.535
				(6.518)	*(10.564)*	*(5.975)*
× initial score				-0.019	0.079	0.147
				(0.105)	*(0.160)*	*(0.091)*
Standardized testing without external comparison	67.772***	86.860***	88.701***	72.689***	77.183**	116.503***
	(17.139)	*(20.263)*	*(21.396)*	*(26.701)*	*(34.691)*	*(31.505)*
× initial score	-0.776***	-0.989***	-1.026***	-0.756***	-0.921**	-1.378***
	(0.175)	*(0.255)*	*(0.260)*	*(0.273)*	*(0.387)*	*(0.377)*
Internal testing	-13.858	-14.734	-26.214	-14.462	-0.669	-44.234
	(12.216)	*(15.155)*	*(17.261)*	*(21.562)*	*(35.177)*	*(33.433)*
× initial score	0.161	0.289**	0.082	0.159	0.087	0.219
	(0.100)	*(0.143)*	*(0.185)*	*(0.201)*	*(0.324)*	*(0.337)*
Internal teacher monitoring	10.432	18.210	-22.463	-0.620	2.077	-42.345
	(25.005)	*(25.338)*	*(32.946)*	*(32.969)*	*(42.956)*	*(43.058)*
× initial score	-0.478*	-0.407	0.077	-0.290	-0.191	0.421
	(0.249)	*(0.289)*	*(0.317)*	*(0.355)*	*(0.506)*	*(0.436)*
Control variables	Yes	Yes	Yes	Yes	Yes	Yes
Country fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Student observations	2,094,856	2,094,705	2,187,415	1,672,041	1,671,914	1,751,351
Country observations	59	59	59	42	42	42
Country-by-wave observations	303	303	303	230	230	230
R^2	0.393	0.349	0.359	0.350	0.316	0.323

Notes: Dependent variable: PISA test score in subject indicated in the header. Least squares regression weighted by students' sampling probability, including country and year fixed effects. Student assessment measures aggregated to the country level. Initial score: country's PISA score in the initial year (centered at 400, so that main-effect coefficient shows effect of assessments on test scores in a country with 400 PISA points in 2000). Sample: student-level observations in six PISA waves 2000-2015. See Table 2 for included control variables. Complete model of specification in column 1 displayed in Table A1. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent.

Table 6: Placebo test with leads of testing reforms

	Math	Science	Reading
	(1)	(2)	(3)
Standardized testing with external comparison	25.104***	24.567***	27.787***
	(6.316)	*(5.242)*	*(7.501)*
Standardized testing without external comparison	-16.172	-3.734	4.660
	(18.139)	*(19.288)*	*(18.490)*
Internal testing	14.305	19.522	-17.675
	(15.367)	*(21.238)*	*(20.325)*
Internal teacher monitoring	-35.785	-38.797*	-31.560
	(22.833)	*(19.796)*	*(19.079)*
Lead (Standardized testing with external comparison)	12.119	4.475	5.746
	(11.045)	*(8.506)*	*(9.351)*
Lead (Standardized testing without external comparison)	-15.195	-11.138	-17.220
	(13.881)	*(16.216)*	*(19.718)*
Lead (Internal testing)	6.965	-7.014	5.567
	(14.408)	*(15.286)*	*(14.069)*
Lead (Internal teacher monitoring)	-5.394	20.922	-15.352
	(17.088)	*(18.269)*	*(17.759)*
Control variables	Yes	Yes	Yes
Country fixed effects	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes
Student observations	1,638,149	1,638,084	1,710,196
Country observations	59	59	59
Country-by-wave observations	235	235	235
R^2	0.396	0.350	0.361

Notes: Dependent variable: PISA test score in subject indicated in the header. Lead indicates values of testing category from subsequent period, i.e., before its later introduction. Least squares regression weighted by students' sampling probability, including country and year fixed effects. Student assessment measures aggregated to the country level. Sample: student-level observations in six PISA waves 2000-2015. See Table 2 for included control variables. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent.

Table 7: Specification tests: Base specification

	No teacher controls	No controls	Long difference (2000+2015 only)
	(1)	(2)	(3)
Standardized testing with external comparison	28.429***	29.902***	61.184***
	(6.067)	*(6.619)*	*(9.981)*
Standardized testing without external comparison	-4.271	0.218	-16.515
	(14.502)	*(13.187)*	*(19.191)*
Internal testing	10.776	13.052	19.131
	(12.001)	*(10.514)*	*(26.395)*
Internal teacher monitoring	-42.255***	-30.877*	-13.438
	(15.604)	*(16.250)*	*(23.881)*
Teacher control variables	No	No	Yes
Other control variables	Yes	No	Yes
Country fixed effects	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes
Student observations	2,094,856	2,094,856	404,344
Country observations	59	59	38
Country-by-wave observations	303	303	76
R^2	0.390	0.256	0.365

Notes: Dependent variable: PISA math test score. Least squares regression weighted by students' sampling probability, including country and year fixed effects. Student assessment measures aggregated to the country level. Sample: student-level observations in six PISA waves 2000-2015. See Table 2 for included control variables. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent.

Table 8: Specification tests: Interacted specification

	No teacher controls	No controls	Long difference (2000+2015 only)		Interactions with four quartiles of initial score			
					×Q1	×Q2	×Q3	×Q4
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Standardized testing with external comparison	37.340***	53.124***	18.944	69.060***	55.899***	26.505***	9.208	18.278
	(5.986)	*(11.586)*	*(24.016)*	*(17.063)*	*(16.514)*	*(7.515)*	*(11.065)*	*(13.847)*
× initial score	-0.249***	-0.440***	0.211	-0.272				
	(0.080)	*(0.144)*	*(0.222)*	*(0.187)*				
Standardized testing without external comparison	74.378***	54.154***	42.848		60.373**	31.831*	-15.650	-67.691**
	(18.061)	*(17.107)*	*(31.020)*		*(26.276)*	*(17.614)*	*(18.383)*	*(27.897)*
× initial score	-0.845***	-0.525***	-0.510					
	(0.183)	*(0.166)*	*(0.335)*					
Internal testing	-10.574	-13.016	-106.185**		-25.596	-11.618	0.771	19.721
	(12.230)	*(14.113)*	*(45.672)*		*(21.609)*	*(13.145)*	*(12.970)*	*(15.521)*
× initial score	0.157	0.166	1.119**					
	(0.097)	*(0.121)*	*(0.473)*					
Internal teacher monitoring	-0.187	-1.592	72.304		55.611	-39.794***	-18.496	-57.127***
	(24.352)	*(30.817)*	*(52.716)*		*(40.507)*	*(14.249)*	*(25.776)*	*(20.785)*
× initial score	-0.411*	-0.255	-1.106*					
	(0.245)	*(0.297)*	*(0.551)*					
Teacher control variables	No	No	Yes	Yes	Yes			
Other control variables	Yes	No	Yes	Yes	Yes			
Country fixed effects	Yes	Yes	Yes	Yes	Yes			
Year fixed effects	Yes	Yes	Yes	Yes	Yes			
Student observations	2,094,856	2,094,856	404,344	404,344	2,094,856			
Country observations	59	59	38	38	59			
Country-by-wave observations	303	303	76	76	303			
R^2	0.392	0.258	0.367	0.365	0.393			

Notes: Dependent variable: PISA math test score. Least squares regression weighted by students' sampling probability, including country and year fixed effects. Student assessment measures aggregated to the country level. Initial score: country's PISA score in the initial year (centered at 400, so that main-effect coefficient shows effect of assessments on test scores in a country with 400 PISA points in 2000). Model in columns (5)-(8) is estimated as one joined model that interacts each assessment measure with four dummies for the quartiles of initial country scores. Sample: student-level observations in six PISA waves 2000-2015. See Table 2 for included control variables. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent.

Table 9: Robustness tests: Base specification

	OECD countries	Non-OECD countries	Control for exclusion rates	Without 2015	Rescaled test scale
	(1)	(2)	(3)	(4)	(5)
Standardized testing with external comparison	29.303***	16.429*	27.431***	31.205***	33.247***
	(7.471)	*(8.387)*	*(6.160)*	*(5.996)*	*(8.937)*
Standardized testing without external comparison	4.671	-10.835	-5.817	-10.664	-10.906
	(15.292)	*(19.542)*	*(13.900)*	*(15.272)*	*(15.499)*
Internal testing	1.727	15.001	5.665	6.381	5.434
	(13.704)	*(14.846)*	*(10.619)*	*(16.582)*	*(9.393)*
Internal teacher monitoring	-25.693	-22.625	-35.308**	-46.460**	-29.108
	(16.190)	*(21.114)*	*(15.460)*	*(20.489)*	*(21.312)*
Control variables	Yes	Yes	Yes	Yes	Yes
Country fixed effects	Yes	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes	Yes
Student observations	1,434,355	660,501	2,045,454	1,679,250	1,698,971
Country observations	35	24	59	59	58
Country-by-wave observations	197	106	289	247	223
R^2	0.283	0.441	0.388	0.399	n.a.

Notes: Dependent variable: PISA math test score. Least squares regression weighted by students' sampling probability, including country and year fixed effects. Student assessment measures aggregated to the country level. Sample: student-level observations in six PISA waves 2000-2015. Rescaled test scale available for waves 2006-2015 only. See Table 2 for included control variables. Robust standard errors adjusted for clustering at the country level in parentheses. Significance level: *** 1 percent, ** 5 percent, * 10 percent.